Predicting Student Emotions in Computer-Human Tutoring Dialogues

Diane J. Litman
University of Pittsburgh
Department of Computer Science
Learning Research and Development Center
Pittsburgh PA, 15260, USA
litman@cs.pitt.edu

Kate Forbes-Riley
University of Pittsburgh
Learning Research and Development Center
Pittsburgh PA, 15260, USA
forbesk@pitt.edu

Abstract

We examine the utility of speech and lexical features for predicting student emotions in computer-human spoken tutoring dialogues. We first annotate student turns for negative, neutral, positive and mixed emotions. We then extract acoustic-prosodic features from the speech signal, and lexical items from the transcribed or recognized speech. We compare the results of machine learning experiments using these features alone or in combination to predict various categorizations of the annotated student emotions. Our best results yield a 19-36% relative improvement in error reduction over a baseline. Finally, we compare our results with emotion prediction in human-human tutoring dialogues.

1 Introduction

This paper explores the feasibility of automatically predicting student emotional states in a corpus of computer-human spoken tutoring dialogues. Intelligent tutoring *dialogue systems* have become more prevalent in recent years (Aleven and Rose, 2003), as one method of improving the performance gap between computer and human tutors; recent experiments with such systems (e.g., (Graesser et al., 2002)) are starting to yield promising empirical results. Another method for closing this performance gap has been to incorporate *affective reasoning* into computer tutoring systems, independently of whether or not the tutor is dialogue-based (Conati et al., 2003; Kort et al., 2001; Bhatt et al., 2004). For example, (Aist et al., 2002) have shown that adding human-provided emotional scaffolding to an automated reading tutor increases student persistence. Our long-term goal is to merge these lines of dialogue and affective tutoring research, by enhancing our intelligent tutoring spoken dialogue system to automatically predict and adapt to student emotions, and to investigate whether this improves learning and other measures of performance.

Previous spoken dialogue research has shown that predictive models of emotion distinctions (e.g., emotional vs. non-emotional, negative vs. non-negative) can be developed using features typically available to a spoken dialogue system in real-time (e.g, acoustic-prosodic, lexical, dialogue, and/or contextual) (Batliner et al., 2000; Lee et al., 2001; Lee et al., 2002; Ang et al., 2002; Batliner et al., 2003; Shafran et al., 2003). In prior work we built on and generalized such research, by defining a three-way distinction between negative, neutral, and positive student emotional states that could be reliably annotated and accurately predicted in *human*-human spoken tutoring dialogues (Forbes-Riley and Litman, 2004; Litman and Forbes-Riley, 2004). Like the non-tutoring studies, our results showed that combining feature types yielded the highest predictive accuracy.

In this paper we investigate the application of our approach to a comparable corpus of *computer*-human tutoring dialogues, which displays many different characteristics, such as shorter utterances, little student initiative, and non-overlapping speech. We investigate whether we can annotate and predict student emotions as accurately and whether the relative utility of speech and lexical features as predictors is the same, especially when the output of the speech recognizer is used (rather than a human transcription of the student speech). Our best models for predicting three different types of emotion classifications achieve accuracies of 66-73%, representing relative improvements of 19-36% over majority class baseline errors. Our computer-human results also show interesting differences compared with comparable analyses of human-human data. Our results provide an empirical basis for enhancing our spoken dialogue tutoring system to automatically predict and adapt to a student model that includes emotional states.

2 Computer-Human Dialogue Data

Our data consists of student dialogues with ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman and Silliman, 2004), a spoken dialogue tutor built on top of the Why2-Atlas concep-

tual physics text-based tutoring system (VanLehn et al., 2002). In ITSPOKE, a student first types an essay answering a qualitative physics problem. ITSPOKE then analyzes the essay and engages the student in spoken dialogue to correct misconceptions and to elicit complete explanations.

First, the Why2-Atlas back-end parses the student essay into propositional representations, in order to find useful dialogue topics. It uses 3 different approaches (symbolic, statistical and hybrid) competitively to create a representation for each sentence, then resolves temporal and nominal anaphora and constructs proofs using abductive reasoning (Jordan et al., 2004). During the dialogue, student speech is digitized from microphone input and sent to the Sphinx2 recognizer, whose stochastic language models have a vocabulary of 1240 words and are trained with 7720 student utterances from evaluations of Why2-Atlas and from pilot studies of ITSPOKE. Sphinx2's best "transcription" (recognition output) is then sent to the Why2-Atlas back-end for syntactic, semantic and dialogue analysis. Finally, the text response produced by Why2-Atlas is sent to the Cepstral text-to-speech system and played to the student. After the dialogue, the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision.

Our corpus of dialogues with ITSPOKE was collected from November 2003 - April 2004, as part of an evaluation comparing ITSPOKE, Why2-Atlas, and human tutoring (Litman et al., 2004). Subjects are University of Pittsburgh students who have never taken college physics, and who are native English speakers. Subjects first read a small document of background physics material, then work through 5 problems (dialogues) with ITSPOKE. The corpus contains 100 dialogues (physics problems) from 20 subjects, with a total of 2445 student turns and 398 unique words. 15 dialogues have been annotated for emotion as described in Section 3. On average, our dialogues last 19.4 minutes and contain 25 student turns. While ITSPOKE's word error rate on this corpus is 31.2%, semantic accuracy is more useful for dialogue evaluation as it does not penalize for unimportant word errors. Semantic analysis based on speech recognition is the same as based on perfect transcription 92.4% of the time. An emotion-annotated corpus example is shown in Figure 1.

3 Annotating Student Turns

In our data, student "emotions"[1] can only be identified indirectly: via what is said and/or how it is said. In (Litman and Forbes-Riley, 2004), we discuss a scheme for manually annotating student *turns* in a *human*-human tutoring dialogue corpus for *intuitively perceived* emotions.[2] These emotions are viewed along a linear scale, shown and defined as follows: **negative** $\longleftarrow$ **neutral** $\longrightarrow$ **positive**.

. . . dialogue excerpt at 18.3 min. into session. . .

ITSPOKE$_{21}$: What is the magnitude of the acceleration of the packet in the horizontal direction?

STUDENT$_{18}$: i don't know (*NEGATIVE*)

ITSPOKE$_{22}$: Ok, let's consider how forces in the vertical direction affect the motion of a body in a horizontal direction in a different scenario. Say an apple falls from a tree. What force(s) are acting on the apple as it falls?

STUDENT$_{19}$: gravity (*NEGATIVE*)

ITSPOKE$_{23}$: Yeah. Are there any other forces acting on the apple as it falls?

STUDENT$_{20}$: no (*NEGATIVE*)

ITSPOKE$_{24}$: Fine. The only force acting on the apple as it falls to the earth is the earth's gravitational force. In which direction does gravitational force act?

STUDENT$_{21}$: down (*POSITIVE*)

ITSPOKE$_{25}$: Yep. Therefore, what is the magnitude of this gravitational force in the horizontal direction?

STUDENT$_{22}$: in the direction of the airplane (*NEUTRAL*)

Figure 1: Annotated Spoken Dialogue Excerpt

Negative: a student turn that expresses emotions such as *confused, bored, irritated*. Evidence of a negative emotion can come from many knowledge sources such as lexical items (e.g., "I don't know" in **student**$_{18}$ in Figure 1), and/or acoustic-prosodic features (e.g., prior-turn pausing in **student**$_{18-20}$).

Positive: a student turn expressing emotions such as *confident, enthusiastic*. An example is **student**$_{21}$, which displays louder speech and faster tempo.

Neutral: a student turn not expressing a negative or positive emotion. An example is **student**$_{22}$, where evidence comes from moderate loudness, pitch and tempo.

We also distinguish **Mixed**: a student turn expressing both positive and negative emotions.

To avoid influencing the annotator's intuitive understanding of emotion expression, and because particular emotional cues are not used consistently

[1] We use the term "emotion" loosely to cover both affects and attitudes that can impact student learning.

[2] Weak and strong expressions of emotions are annotated.

or unambiguously across speakers, our annotation manual does *not* associate particular cues with particular emotion labels. Instead, it contains examples of labeled dialogue excerpts (as in Figure 1, except on human-human data) with links to corresponding audio files. The cues mentioned in the discussion of Figure 1 above were elicited during *post*-annotation discussion of the emotions, and are presented here for expository use only. (Litman and Forbes-Riley, 2004) further details our annotation scheme and discusses how it builds on related work.

To analyze the reliability of the scheme on our new *computer*-human data, we selected 15 transcribed dialogues from the corpus described in Section 2, yielding a dataset of 333 student turns, where approximately 30 turns came from each of 10 subjects. The 333 turns were separately annotated by two annotators following the emotion annotation scheme described above.

We focus here on three analyses of this data, itemized below. While the first analysis provides the most fine-grained distinctions for triggering system adaptation, the second and third (simplified) analyses correspond to those used in (Lee et al., 2001) and (Batliner et al., 2000), respectively. These represent alternative potentially useful triggering mechanisms, and are worth exploring as they might be easier to annotate and/or predict.

- **Negative, Neutral, Positive (NPN)**: *mixeds* are conflated with **neutrals**.
- **Negative, Non-Negative (NnN)**: *positives, mixeds, neutrals* are conflated as **non-negatives**.
- **Emotional, Non-Emotional (EnE)**: *negatives, positives, mixeds* are conflated as **Emotional**; *neutrals* are **Non-Emotional**.

Tables 1-3 provide a confusion matrix for each analysis summarizing inter-annotator agreement. The rows correspond to the labels assigned by annotator 1, and the columns correspond to the labels assigned by annotator 2. For example, the annotators agreed on 89 negatives in Table 1.

In the NnN analysis, the two annotators agreed on the annotations of 259/333 turns achieving 77.8% agreement, with Kappa = 0.5. In the EnE analysis, the two annotators agreed on the annotations of 220/333 turns achieving 66.1% agreement, with Kappa = 0.3. In the NPN analysis, the two annotators agreed on the annotations of 202/333 turns achieving 60.7% agreement, with Kappa = 0.4. This inter-annotator agreement is on par with that of prior studies of emotion annotation in naturally occurring computer-human dialogues (e.g., agreement of 71% and Kappa of 0.47 in (Ang et al., 2002), Kappa of 0.45 and 0.48 in (Narayanan, 2002), and Kappa ranging between 0.32 and 0.42 in (Shafran et al., 2003)). A number of researchers have accommodated for this low agreement by exploring ways of achieving consensus between disagreed annotations, to yield 100% agreement (e.g (Ang et al., 2002; Devillers et al., 2003)). As in (Ang et al., 2002), we will experiment below with predicting emotions using both our agreed data and consensus-labeled data.

	negative	**non-negative**
negative	**89**	36
non-negative	38	**170**

Table 1: NnN Analysis Confusion Matrix

	emotional	**non-emotional**
emotional	**129**	43
non-emotional	70	**91**

Table 2: EnE Analysis Confusion Matrix

	negative	**neutral**	**positive**
negative	**89**	30	6
neutral	32	**94**	38
positive	6	19	**19**

Table 3: NPN Analysis Confusion Matrix

4 Extracting Features from Turns

For each of the 333 student turns described above, we next extracted the set of features itemized in Figure 2, for use in the machine learning experiments described in Section 5.

Motivated by previous studies of emotion prediction in spontaneous dialogues (Ang et al., 2002; Lee et al., 2001; Batliner et al., 2003), our acoustic-prosodic features represent knowledge of pitch, energy, duration, tempo and pausing. We further restrict our features to those that can be computed automatically and in real-time, since our goal is to use such features to trigger online adaptation in ITSPOKE based on predicted student emotions. F0 and RMS values, representing measures of pitch and loudness, respectively, are computed using Entropic Research Laboratory's pitch tracker, *get_f0*, with no post-correction. Amount of Silence is approximated as the proportion of zero f0 frames for the turn. Turn Duration and Prior Pause Duration are computed

Acoustic-Prosodic Features

- 4 fundamental frequency (f0): max, min, mean, standard deviation
- 4 energy (RMS): max, min, mean, standard deviation
- 4 temporal: amount of silence in turn, turn duration, duration of pause prior to turn, speaking rate

Lexical Features

- human-transcribed lexical items in the turn
- ITSPOKE-recognized lexical items in the turn

Identifier Features: subject, gender, problem

Figure 2: Features Per Student Turn

automatically via the start and end turn boundaries in ITSPOKE logs. Speaking Rate is automatically calculated as #syllables per second in the turn.

While acoustic-prosodic features address *how* something is said, lexical features representing *what* is said have also been shown to be useful for predicting emotion in spontaneous dialogues (Lee et al., 2002; Ang et al., 2002; Batliner et al., 2003; Devillers et al., 2003; Shafran et al., 2003). Our first set of lexical features represents the *human* transcription of each student turn as a word occurrence vector (indicating the lexical items that are present in the turn). This feature represents the "ideal" performance of ITSPOKE with respect to speech recognition. The second set represents ITSPOKE's actual best speech recognition hypothesis of what is said in each student turn, again as a word occurrence vector.

Finally, we recorded for each turn the 3 "identifier" features shown last in Figure 2. Prior studies (Oudeyer, 2002; Lee et al., 2002) have shown that "subject" and "gender" can play an important role in emotion recognition. "Subject" and "problem" are particularly important in our tutoring domain because students will use our system repeatedly, and problems are repeated across students.

5 Predicting Student Emotions

5.1 Feature Sets and Method

We next created the 10 feature sets in Figure 3, to study the effects that various feature combinations had on predicting emotion. We compare an acoustic-prosodic feature set ("sp"), a human-transcribed lexical items feature set ("lex") and an ITSPOKE-recognized lexical items feature set ("asr"). We further compare feature sets combining acoustic-prosodic and either transcribed or recognized lexical items ("sp+lex", "sp+asr"). Finally, we compare each of these 5 feature sets with an identical set supplemented with our 3 identifier features ("+id").

sp: 12 acoustic-prosodic features
lex: human-transcribed lexical items
asr: ITSPOKE recognized lexical items
sp+lex: combined sp and lex features
sp+asr: combined sp and asr features
+id: each above set + 3 identifier features

Figure 3: Feature Sets for Machine Learning

We use the Weka machine learning software (Witten and Frank, 1999) to automatically learn our emotion prediction models. In our human-human dialogue studies (Litman and Forbes, 2003), the use of boosted decision trees yielded the most robust performance across feature sets so we will continue their use here.

5.2 Predicting Agreed Turns

As in (Shafran et al., 2003; Lee et al., 2001), our first study looks at the clearer cases of emotional turns, i.e. only those student turns where the two annotators agreed on an emotion label.

Tables 4-6 show, for each emotion classification, the mean accuracy (%correct) and standard error (SE) for our 10 feature sets (Figure 3), computed across 10 runs of 10-fold cross-validation.[3] For comparison, the accuracy of a standard baseline algorithm (MAJ), which always predicts the majority class, is shown in each caption. For example, Table 4's caption shows that for NnN, always predicting the majority class of non-negative yields an accuracy of 65.65%. In each table, the accuracies are labeled for how they compare statistically to the relevant baseline accuracy (w = worse, s = same, b = better), as automatically computed in Weka using a two-tailed t-test ($p < .05$).

First note that almost every feature set significantly outperforms the majority class baseline, across all emotion classifications; the only exceptions are the speech-only feature sets without identifier features ("sp-id") in the NnN and EnE tables, which perform the same as the baseline. These results suggest that without any subject or task specific information, acoustic-prosodic features alone

[3] For each cross-validation, the training and test data are drawn from utterances produced by the same set of speakers. A separate experiment showed that testing on one speaker and training on the others, averaged across all speakers, does not significantly change the results.

are not useful predictors for our two binary classification tasks, at least in our computer-human dialogue corpus. As will be discussed in Section 6, however, "sp-id" feature sets are useful predictors in human-human tutoring dialogues.

Feat. Set	-id	SE	+id	SE
sp	64.10s	0.80	70.66b	0.76
lex	68.20b	0.41	72.74b	0.58
asr	**72.30**b	0.58	70.51b	0.59
sp+lex	71.78b	0.77	72.43b	0.87
sp+asr	69.90b	0.57	**71.44b**	0.68

Table 4: %Correct, NnN Agreed, MAJ (non-negative) = 65.65%

Feat. Set	-id	SE	+id	SE
sp	59.18s	0.75	70.68b	0.89
lex	63.18b	0.82	75.64b	0.37
asr	**66.36**b	0.54	**72.91**b	0.35
sp+lex	63.86b	0.97	69.59b	0.48
sp+asr	65.14b	0.82	69.64b	0.57

Table 5: %Correct, EnE Agreed, MAJ (emotional) = 58.64%

Feat. Set	-id	SE	+id	SE
sp	55.49b	1.01	62.03b	0.91
lex	52.66b	0.62	67.84b	0.66
asr	57.95b	0.67	**65.70**b	0.50
sp+lex	62.08b	0.56	63.52b	0.48
sp+asr	**61.22**b	1.20	62.23b	0.86

Table 6: %Correct, NPN Agreed, MAJ (neutral) = 46.52%

Further note that adding identifier features to the "-id" feature sets almost always improves performance, although this difference is not always significant[4]; across tables the "+id" feature sets outperform their "-id" counterparts across all feature sets and emotion classifications except one (NnN "asr"). Surprisingly, while (Lee et al., 2002) found it useful to develop separate gender-based emotion prediction models, in our experiment, gender is the only identifier that does not appear in any learned model. Also note that with the addition of identifier features, the speech-only feature sets (sp+id) now do outperform the majority class baselines for all three emotion classifications.

[4]For any feature set, the mean +/- 2*SE = the 95% confidence interval. If the confidence intervals for two feature sets are non-overlapping, then their mean accuracies are significantly different with 95% confidence.

With respect to the relative utility of lexical versus acoustic-prosodic features, without identifier features, using only lexical features ("lex" or "asr") almost always produces statistically better performance than using only speech features ("sp"); the only exception is NPN "lex", which performs statistically the same as NPN "sp". This is consistent with others' findings, e.g., (Lee et al., 2002; Shafran et al., 2003). When identifier features are added to both, the lexical sets don't always significantly outperform the speech set; only in NPN and EnE "lex+id" is this the case. For NnN, just as using "sp+id" rather than "sp-id" improved performance when compared to the majority baseline, the addition of the identifier features also improves the utility of the speech features when compared to the lexical features.

Interestingly, although we hypothesized that the "lex" feature sets would present an upper bound on the performance of the "asr" sets, because the human transcription is more accurate than the speech recognizer, we see that this is not consistently the case. In fact, in the "-id" sets, "asr" always significantly outperforms "lex". A comparison of the decision trees produced in either case, however, does not reveal why this is the case; words chosen as predictors are not very intuitive in either case (e.g., for NnN, an example path through the learned "lex" decision tree says predict *negative* if the utterance contains the word *will* but does not contain the word *decrease*). Understanding this result is an area for future research. Within the "+id" sets, we see that "lex" and "asr" perform the same in the NnN and NPN classifications; in EnE "lex+id" significantly outperforms "asr+id". The utility of the "lex" features compared to "asr" also increases when combined with the "sp" features (with and without identifiers), for both NnN and NPN.

Moreover, based on results in (Lee et al., 2002; Ang et al., 2002; Forbes-Riley and Litman, 2004), we hypothesized that combining speech and lexical features would result in better performance than either feature set alone. We instead found that the relative performance of these sets depends both on the emotion classification being predicted and the presence or absence of "id" features. Although consistently with prior research we find that the combined feature sets usually outperform the speech-only feature sets, the combined feature sets frequently perform worse than the lexical-only feature sets. However, we will see in Section 6 that combining knowledge sources does improve prediction performance in human-human dialogues.

Finally, the bolded accuracies in each table sum-

marize the best-performing feature sets with and without identifiers, with respect to both the %Corr figures shown in the tables, as well as to relative improvement in error reduction over the baseline (MAJ) error[5], after *excluding* all the feature sets containing "lex" features. In this way we give a better estimate of the best performance our system could accomplish, given the features it can currently access from among those discussed. These best-performing feature sets yield relative improvements over their majority baseline errors ranging from 19-36%. Moreover, although the NPN classification yields the lowest raw accuracies, it yields the highest relative improvement over its baseline.

5.3 Predicting Consensus Turns

Following (Ang et al., 2002; Devillers et al., 2003), we also explored *consensus labeling*, both with the goal of increasing our usable data set for prediction, and to include the more difficult annotation cases. For our consensus labeling, the original annotators revisited each originally disagreed case, and through discussion, sought a consensus label. Due to consensus labeling, agreement rose across all three emotion classifications to 100%. Tables 7-9 show, for each emotion classification, the mean accuracy (%correct) and standard error (SE) for our 10 feature sets.

Feat. Set	-id	SE	+id	SE
sp	59.10w	0.57	64.20b	0.52
lex	63.70s	0.47	68.64b	0.41
asr	**66.26**b	0.71	**68.13**b	0.56
sp+lex	64.69b	0.61	65.40b	0.63
sp+asr	65.99b	0.51	67.55b	0.48

Table 7: %Corr., NnN Consensus, MAJ=62.47%

Feat. Set	-id	SE	+id	SE
sp	56.13s	0.94	59.30b	0.48
lex	52.07w	0.34	65.37b	0.47
asr	53.78s	0.66	**64.13**b	0.51
sp+lex	60.96b	0.76	63.01b	0.62
sp+asr	**57.84**s	0.73	60.89b	0.38

Table 8: %Corr., EnE Consensus, MAJ=55.86%

A comparison with Tables 4-6 shows that overall, using consensus-labeled data decreased the performance across all feature sets and emotion classifications. This was also found in (Ang et al., 2002). Moreover, it is no longer the case that every feature set performs as well as or better than their baselines[6]; within the "-id" sets, NnN "sp" and EnE "lex" perform significantly worse than their baselines. However, again we see that the "+id" sets do consistently better than the "-id" sets and moreover always outperform the baselines.

Feat. Set	-id	SE	+id	SE
sp	48.97s	0.66	51.90b	0.40
lex	47.86s	0.54	57.28b	0.44
asr	51.09b	0.66	53.41b	0.66
sp+lex	53.41b	0.62	54.20b	0.86
sp+asr	**52.50**b	0.42	**53.84**b	0.42

Table 9: %Corr., NPN Consensus, MAJ=48.35%

We also see again that using only lexical features almost always yields better performance than using only speech features. In addition, we again see that the "lex" feature sets perform comparably to the "asr" feature sets, rather than outperforming them as we first hypothesized. And finally, we see again that while in most cases combining speech and lexical features yields better performance than using only speech features, the combined feature sets in most cases perform the same or worse than the lexical feature sets. As above, the bolded accuracies summarize the best-performing feature sets from each emotion classification, after *excluding* all the feature sets containing "lex" to give a better estimate of actual system performance. The best-performing feature sets in the consensus data yield an 11%-19% relative improvement in error reduction compared to the majority class prediction, which is a lower error reduction than seen for agreed data. Moreover, the NPN classification yields the lowest accuracies and the lowest improvements over its baseline.

6 Comparison with Human Tutoring

While building ITSPOKE, we collected a corresponding corpus of spoken human tutoring dialogues, using the same experimental methodology as for our computer tutoring corpus (e.g. same subject pool, physics problems, web and audio interface, etc); the only difference between the two corpora is whether the tutor is human or computer. As discussed in (Forbes-Riley and Litman, 2004), two annotators had previously labeled 453 turns in this corpus with the emotion annotation scheme discussed in Section 3, and performed a preliminary set of machine learning experiments (different from those reported above). Here, we perform the exper-

[5]Relative improvement over the baseline (MAJ) error for feature set x = $\frac{error(baseline)-error(x)}{error(baseline)}$, where error(x) is 100 minus the %Corr(x) value shown in Tables 4-6.

[6]The majority class for EnE Consensus is non-emotional; all others are unchanged.

	NnN				EnE				NPN			
FS	**-id**	**SE**	**+id**	**SE**	**-id**	**SE**	**+id**	**SE**	**-id**	**SE**	**+id**	**SE**
sp	77.46	0.42	77.56	0.30	84.71	0.39	84.66	0.40	73.09	0.68	74.18	0.40
lex	80.74	0.42	80.60	0.34	88.86	0.26	86.23	0.34	78.56	0.45	77.18	0.43
sp+lex	81.37	0.33	80.79	0.41	87.74	0.36	88.31	0.29	79.06	0.38	78.03	0.33

Table 10: Human-Human %Correct, NnN MAJ=72.21%; EnE MAJ=50.86%; NPN MAJ=53.24%

iments from Section 5.2 on this annotated human tutoring data, as a step towards understand the differences between annotating and predicting emotion in human versus computer tutoring dialogues.

With respect to inter-annotator agreement, in the NnN analysis, the two annotators had 88.96% agreement (Kappa = 0.74). In the EnE analysis, the annotators had 77.26% agreement (Kappa = 0.55). In the NPN analysis, the annotators had 75.06% agreement (Kappa = 0.60). A comparison with the results in Section 3 shows that all of these figures are higher than their computer tutoring counterparts.

With respect to predictive accuracy, Table 10 shows our results for the agreed data. A comparison with Tables 4-6 shows that overall, the human-human data yields increased performance across all feature sets and emotion classifications, although it should be noted that the human-human corpus is over 100 turns larger than the computer-human corpus. Every feature set performs significantly better than their baselines. However, unlike the computer-human data, we don't see the "+id" sets performing better than the "-id" sets; rather, both sets perform about the same. We do see again the "lex" sets yielding better performance than the "sp" sets. However, we now see that in 5 out of 6 cases, combining speech and lexical features yields better performance than using either "sp" or "lex" alone. Finally, these feature sets yield a relative error reduction of 42.45%-77.33% compared to the majority class predictions, which is far better than in our computer tutoring experiments. Moreover, the EnE classification yields the highest raw accuracies and relative improvements over baseline error.

We hypothesize that such differences arise in part due to differences between the two corpora: 1) student turns with the computer tutor are much shorter than with the human tutor (and thus contain less emotional content - making both annotation and prediction more difficult), 2) students respond to the computer tutor differently and perhaps more idiosyncratically than to the human tutor, 3) the computer tutor is less "flexible" than the human tutor (allowing little student initiative, questions, groundings, contextual references, etc.), which also effects student emotional response and its expression.

7 Conclusions and Current Directions

Our results show that acoustic-prosodic and lexical features can be used to automatically predict student emotion in computer-human tutoring dialogues. We examined emotion prediction using a classification scheme developed for our prior human-human tutoring studies (negative/positive/neutral), as well as using two simpler schemes proposed by other dialogue researchers (negative/non-negative, emotional/non-emotional). We used machine learning to examine the impact of different feature sets on prediction accuracy. Across schemes, our feature sets outperform a majority baseline, and lexical features outperform acoustic-prosodic features. While adding identifier features typically also improves performance, combining lexical and speech features does not. Our analyses also suggest that prediction in consensus-labeled turns is harder than in agreed turns, and that prediction in our computer-human corpus is harder and based on somewhat different features than in our human-human corpus.

Our continuing work extends this methodology with the goal of enhancing ITSPOKE to predict and adapt to student emotions. We continue to manually annotate ITSPOKE data, and are exploring partial automation via semi-supervised machine learning (Maeireizo-Tokeshi et al., 2004). Further manual annotation might also improve reliability, as understanding systematic disagreements can lead to coding manual revisions. We are also expanding our feature set to include features suggested in prior dialogue research, tutoring-dependent features (e.g., pedagogical goal), and other features available in our logs (e.g., semantic analysis). Finally, we will explore how the recognized emotions can be used to improve system performance. First, we will label *human* tutor adaptations to emotional student turns in our human tutoring corpus; this labeling will be used to formulate adaptive strategies for ITSPOKE, and to determine which of our three prediction tasks best triggers adaptation.

Acknowledgments

This research is supported by NSF Grants 9720359 & 0328431. Thanks to the Why2-Atlas team and S. Silliman for system design and data collection.

References

G. Aist, B. Kort, R. Reilly, 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. Intelligent Tutoring Systems*.

V. Aleven and C. P. Rose, editors. 2003. *Proc. AI in Education Workshop on Tutorial Dialogue Systems: With a View toward the Classroom*.

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. International Conf. on Spoken Language Processing (ICSLP)*.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 2000. Desperately seeking emotions: Actors, wizards, and human beings. In *Proc. ISCA Workshop on Speech and Emotion*.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. How to find trouble in communication. *Speech Communication*, 40:117–143.

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

C. Conati, R. Chabbal, and H. Maclaren. 2003. A study on using biometric sensors for monitoring user emotions in educational games. In *Proc. User Modeling Workshop on Assessing and Adapting to User Attitudes and Effect: Why, When, and How?*

L. Devillers, L. Lamel, and I. Vasilescu. 2003. Emotion detection in task-oriented spoken dialogs. In *Proc. IEEE International Conference on Multimedia & Expo (ICME)*.

K. Forbes-Riley and D. Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proc. Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL)*.

A. Graesser, K. VanLehn, C. Rose, P. Jordan, and D. Harter. 2002. Intelligent tutoring systems with conversational dialogue. *AI Magazine*.

P. W. Jordan, M. Makatchev, and K. VanLehn. 2004. Combining competing language understanding approaches in an intelligent tutoring system. In *Proc. Intelligent Tutoring Systems*.

B. Kort, R. Reilly, and R. W. Picard. 2001. An affective model of interplay between emotions and learning: Reengineering educational pedagogy - building a learning companion. In *International Conf. on Advanced Learning Technologies*.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2001. Recognition of negative emotions from the speech signal. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2002. Combining acoustic and language information for emotion recognition. In *International Conf. on Spoken Language Processing (ICSLP)*.

D. Litman and K. Forbes-Riley. 2004. Annotating student emotional states in spoken tutoring dialogues. In *Proc. 5th SIGdial Workshop on Discourse and Dialogue*.

D. Litman and K. Forbes. 2003. Recognizing emotion from student speech in tutoring dialogues. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

D. Litman and S. Silliman. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf. of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL)*.

D. J. Litman, C. P. Rose, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. 2004. Spoken versus typed human and computer dialogue tutoring. In *Proc. Intelligent Tutoring Systems*.

B. Maeireizo-Tokeshi, D. Litman, and R. Hwa. 2004. Co-training for predicting emotions with spoken dialogue data. In *Companion Proc. Assoc. for Computational Linguistics (ACL)*.

S. Narayanan. 2002. Towards modeling user behavior in human-machine interaction: Effect of errors and emotions. In *Proc. ISLE Workshop on Dialogue Tagging for Multi-modal Human Computer Interaction*.

P-Y. Oudeyer. 2002. The production and recognition of emotions in speech: Features and Algorithms. *International Journal of Human Computer Studies*, 59(1-2):157–183.

I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*.

K. VanLehn, P. W. Jordan, C. P. 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. Intelligent Tutoring Systems*.

I. H. Witten and E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java implementations*.