Annotating Student Emotional States in Spoken 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 present an annotation scheme for student emotions in tutoring dialogues. Analyses of our scheme with respect to interannotator agreement and predictive accuracy indicate that our scheme is reliable in our domain, and that our emotion labels can be predicted with a high degree of accuracy. We discuss issues concerning the implementation of emotion prediction and adaptation in the computer tutoring dialogue system we are developing.

1 Introduction

This paper describes a coding scheme for annotating student emotional states in spoken dialogue tutoring corpora, and analyzes the scheme not only for its reliability, but also for its utility in developing a spoken dialogue tutoring system that can model and respond to student emotions. Motivation for this work comes from the performance discrepancy between human tutors and current machine tutors: typically, students tutored by human tutors achieve higher learning gains than students tutored by computer tutors. The development of computational tutorial *dialogue* systems (Rosé and Aleven, 2002) represents one method of closing this performance gap, e.g. it is hypothesized that dialogue-based tutors allow greater adaptivity to students' beliefs and misconceptions. Another method for closing this performance gap involves incorporating *emotion* prediction and adaptation into computer tutors (Kort et al., 2001; Evens, 2002). For example (Aist et al., 2002) have shown that adding human-provided emotional scaffolding to an automated reading tutor increases student persistence. This suggests that the success of computer *dialogue* tutors could be increased by responding to both *what* a student says and *how* s/he says it, e.g. with *confidence* or *uncertainty*.

To assess the impact of adding emotion modeling to dialogue tutoring systems, we are building ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system), a spoken dialogue system that uses the Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2002) as its "back-end."[1] Our first step towards incorporating emotion processing into ITSPOKE is to develop a reliable annotation scheme for student emotions. Our next step will be to use the data that has been annotated according to this scheme to enhance ITSPOKE to dynamically predict and adapt to student emotions. This adds additional constraints on our annotation scheme besides good reliability, namely that our annotations are predictable by ITSPOKE with a high degree of accuracy (automatically and in real-time), and that they are expressive enough to support the range of desired system adaptations.

In Section 2 we review previous work in emotion annotation for spoken dialogue systems. In Section 3 we discuss our tutoring research project and corpora. In Section 4 we present an emotion annotation scheme for this domain. In Section 5 we analyze our scheme with respect to interannotator agreement and predictive accuracy, using a corpus of human tutoring dialogues. Our agreement indicates that our scheme is reliable, while machine learning experiments on annotated data indicate that our emotion labels can be predicted with a high degree of accuracy. In Section 6 we analyze more expressive versions of our scheme, and discuss differences between annotating human and computer spoken tutoring dialogues.

2 Prior Research on Emotion

Developing a descriptive theory of emotion is a complex research topic, viewed from either a theoretical or an empirical standpoint (Cowie et al., 2001). Some researchers have proposed a variety of "fundamental" human emotions, while others have argued that emotions

[1] We also use ITSPOKE to examine the utility of building *spoken* dialogue tutors (e.g. (Litman and Forbes, 2003)).

are best represented componentially, in terms of multiple dimensions. Despite this lack of a well-defined descriptive framework, there has been great recent interest in predicting emotional states, using information extracted from a person's text, speech, physiology, facial expressions, eye gaze, etc. (Pantic and Rothkrantz, 2003).

In the area of emotional speech, most research has used databases of speech read by actors or native speakers as training data for developing emotion predictors (Holzapfel et al., 2002; Liscombe et al., 2003). In this work the set of emotions to be read is predefined before the utterance is spoken, rather than annotated after the fact. One problem with this approach is that such prototypical emotional speech does not necessarily reflect natural speech (Batliner et al., 2003), e.g. the way one acts an emotion is not necessarily the same as the way one naturally expresses an emotion. Moreover, actors repeatedly reading the same sentence are restricted to conveying different emotions using only acoustic and prosodic features, while in natural interactions a much wider feature variety is available (e.g., lexical, dialogue).

As a result of these problems, researchers motivated by spoken dialogue applications have instead started to train emotion predictors using naturally-occurring speech that has been hand-annotated for various emotions (Ang et al., 2002; Batliner et al., 2003; Lee et al., 2001; Litman and Forbes, 2003). However, this requires researchers to first develop a scheme for annotating emotions in naturally-occurring spoken dialogue corpora. Although emotion annotation of natural corpora (typically at the turn or utterance level) has been addressed in various domains, little has yet been done in the educational setting. Although not yet tested, (Evens, 2002) has hypothesized adaptive strategies; for example, if detecting *frustration*, the system should respond to hedges and self-deprecation, by supplying praise and restructuring the problem. A comparison of our annotation scheme and prior non-tutoring schemes is presented in Section 4.4.

3 The ITSPOKE System and Corpora

In ITSPOKE, a student types an essay answering a qualitative physics problem. The ITSPOKE computer tutor then engages the student in spoken dialogue to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. Student speech is digitized from microphone input and sent to the Sphinx2 recognizer, whose most probable "transcription" output is then sent to the Why2-Atlas back-end for syntactic, semantic and dialogue analysis. The text response produced by Why2-Atlas is sent to the Cepstral text-to-speech system. A formal evaluation of ITSPOKE began in November 2003; to date we have collected 50 dialogues from 10 students. A corpus example is shown in Figure 4, Appendix A. Corpus collection uses the same experimental procedure as our human-human tutoring corpus, described next.

Our Human-Human Spoken Dialogue Tutoring Corpus contains spoken dialogues collected via a web interface supplemented with a high-quality audio link, where the human tutor performs the same task as ITSPOKE. The experimental procedure for collecting both corpora is as follows: 1) students are given a pre-test measuring their physics knowledge, 2) students read through a small document of background material, 3) students use the web and voice interface to work through a set of training problems (dialogues) with the tutor, and 4) students are given a post-test that is similar to the pre-test. Subjects are University of Pittsburgh students who have never taken college physics and who are native English speakers. One tutor currently participates. To date we have collected 149 dialogues from 17 students. Annotated (see Section 4) corpus examples are shown in Figure 1 and Figure 2 (Appendix A) (punctuation added for clarity).

. . . dialogue excerpt at 5.2 minutes into session. . .

TUTOR$_7$: Suppose you apply equal force by pushing them. Then uh what will happen to their motion?

STUDENT$_8$: Um the one that's heavier...uh, the acc- acceleration won't be as great. *(NEGATIVE, UNCERTAIN)*

TUTOR$_8$: The one which is...

STUDENT$_9$: Heavier? *(NEGATIVE, UNCERTAIN)*

TUTOR$_9$: Mm, well, uh, is that your common-

STUDENT$_{10}$: Er I'm sorry- I'm sorry- the one with most mass. You- *(POSITIVE, CONFIDENT)*

TUTOR$_{10}$: (lgh) Yeah, the one with more mass will- if you- if the mass is more and force is the same then which one will accelerate more?

STUDENT$_{11}$: Which one will move more? *(NEGATIVE, CONFUSED)*

TUTOR$_{11}$: Mm, which one will accelerate more?

STUDENT$_{12}$: The- the one with the least amount of mass? *(NEGATIVE, UNCERTAIN)*

TUTOR$_{12}$: Yeah, but what you said was different isn't it? So uh you are applying Newton's law of uh second law of motion: F is equal to M times A. And uh you apply equal force on both the containers, then the one which is less massive will accelerate more.

STUDENT$_{13}$: Right. *(WEAK POSITIVE, CONFIDENT)*

Figure 1: Annotated Excerpt (Human Spoken Corpus)

4 Annotation Scheme

In our spoken dialogue tutoring corpora, student emotional states can only be identified indirectly – via what a student says and/or how s/he says it. Furthermore, such evidence is not always obvious, unambiguous, or consistent. For example, a student may express anger through the use of swear words, or through a particular tone of

voice, or via a combination of signals, or not at all. Moreover, another student may present some of these same signals even when s/he does not feel anger.

Our objective is nevertheless to develop a reliable annotation scheme across annotators, for manually labeling the student *turns* in our spoken tutoring dialogues for *perceived expressions of emotion*.

4.1 Emotion Classes

In our current annotation scheme, perceived expressions of emotion are viewed along a linear scale, as shown and defined below: **negative** $\longleftarrow$ **neutral** $\longrightarrow$ **positive**

Negative: a student turn that strongly expresses emotions such as *confused, bored, irritated, uncertain, sad*. Examples in Figure 1 include $\textbf{student}_8$ and $\textbf{student}_{11}$. Evidence[2] for the negative emotions in these turns includes syntax (constructions such as questions), disfluencies, and acoustic-prosodic features.

Positive: a student turn that strongly expresses emotions such as *confident, enthusiastic*. An example is $\textbf{student}_{10}$ in Figure 1, where evidence of a positive emotion comes primarily from acoustic-prosodic features.

Neutral: a student turn not strongly expressing a negative or positive emotion.

In addition to these three **main** emotion classes, we also distinguish three *minor* emotion classes:

Weak Negative: a student turn that weakly expresses negative emotions.

Weak Positive: a student turn that weakly expresses positive emotions. An example is $\textbf{student}_{13}$ in Figure 1, where evidence is primarily lexical ("right").

Mixed: a student turn that strongly expresses both positive and negative emotions: Case 1) multi-utterance turns where one utterance is judged positive and another, negative. Case 2) turns where the simultaneous strong expression of negative and positive emotions is perceived. Case 2 is often due to conflicting domains (Section 4.2), e.g. boredom with tutoring but confidence about physics.

4.2 Relativity and Domains of Emotion Classes

Our emotion annotation is relative to both context and task. By *context-relative* we mean that a student turn in our tutoring dialogues is identified as expressing emotion relative to the other student turns in that dialogue. By *task-relative* we mean that a student turn perceived during tutoring as expressing an emotion might not be perceived as expressing the same emotion with the same strength in another situation. For example, consider the context of a tutoring session, where a student has been answering tutor questions with apparent ease. If the tutor then asks another question, and the student responds slowly, saying "Um, now I'm confused", this turn would likely be labeled **negative**. However, in the context of a heated argument between two people, this same turn might be labeled as a *weak negative*, or even *weak positive*.

We also annotate emotion with respect to multiple domains. One focus of our annotation scheme is expressions of emotion that pertain to *the physics material being learned* ("PHYS" domain). For example, a student may express confusion or confidence about the physics material. Another focus of our scheme is expressions of emotion that pertain to *the tutoring process*, including attitudes towards the tutor, the dialogue, and/or being tutored ("TUT" domain). For example, a student may express boredom or amusement with the tutoring.

4.3 Specific Annotation Instructions

Our annotation scheme is detailed in an online, audio-enhanced emotion labeling manual. As shown in Figure 3 (Appendix A), the emotion annotation is performed using (our customization of) Wavesurfer, an open source sound visualization and manipulation tool. The "Tutor Speech" and "Student Speech" panes show a portion of the tutor and student speech files, while the "Tutor Text" and "Student Text" show the associated transcriptions, where vertical lines correspond to turn segmentations.[3] There are three additional panes for emotion annotation:

The **EMOa pane** records the annotator's judgment of the expressed emotion class for each turn, e.g. the six emotion classes described in Section 4.1: **negative**, *weak negative*, **neutral**, *weak positive*, **positive**, *mixed*. Annotators are instructed to focus on expressed emotions in the PHYS domain. If an additional expressed emotion in the TUT domain is perceived, this is noted in the NOTES pane (e.g. "amused/TUT"). If no expressed emotion is perceived in the PHYS domain, any expressed emotion in the TUT domain is labeled in the EMOa pane, and noted (e.g. "TUT") in the NOTES pane. Domain indecision is also noted (e.g. "TUT/PHYS?") in the NOTES pane.

The **EMOb pane** further specifies the annotations in the EMOa pane, by recording a specific expressed emotion for each turn. Our current list of specific emotions contains those that we believe will be useful for triggering ITSPOKE adaptation. Specific negative emotions are: *uncertain, confused, sad, bored, irritated*. Specific positive emotions are: *confident, enthusiastic*. Our manual includes glosses for these specific emotions, formulated using synonyms and/or hyponyms that are currently not distinguished. For example, our gloss for *enthusiastic* includes *interested, pleased, amused*. There are also complex labels combining multiple specific emotions within a class (e.g. *uncertain+sad, confident+enthusiastic*). If

[2] Determined in *post*-annotation discussion (see Section 4.4).

[3] Transcription and turn-segmentation of the human-human dialogues were also done within Wavesurfer, by a paid transcriber prior to emotion annotation.

the annotator judges a specific emotion that is not listed (or lacks a close substitute), s/he selects the label *other*, and lists the alternative(s) in the NOTES pane. If the annotator selected **mixed** (case 1) in the EMOa pane, s/he subdivides the turn into utterances in the EMOb pane and provides a specific emotion label for each utterance. If the annotator selected **mixed** (case 2) in the EMOa pane, s/he selects the label *other* in the EMOb pane, and comments on the indecision in the NOTES pane.

The **NOTES pane** records any additional annotator comments concerning their judgment, the annotation, etc.

Because our annotation is student-, context-, and task-specific, our manual first instructs the annotator to listen to each dialogue at least once before annotating, to secure an intuition of how and with what range emotional expression is displayed. S/he is also instructed to not assume that all dialogues will begin with neutral student turns. S/he is however reminded that it is not necessary to assign a non-neutral label to every turn. Finally, s/he is told to ignore correctness when annotating, because a correct answer to a tutor question can express uncertainty, and an incorrect answer can express confidence.

Our manual also describes two default conventions for our annotation scheme, which can however be overridden by the annotator's intuitive judgment and/or other extenuating considerations (e.g. irony, etc), as described below:

1) By definition, a question expresses strong uncertainty or confusion. Thus if a student turn consists only of a question, its default label is **negative**. However:

a) If the turn consists of multiple utterances, one of which is a question, and the other(s) expresses a positive emotion, then the turn should be labeled **mixed** and subdivided (e.g. "What directions are the forces acting in? Gravity is only acting in the down direction").

b) The domain must be considered. For example, defaults in one domain can be overridden if the turn expresses a contrasting emotion in the other domain.

2) Many student turns in our dialogues are very short, containing only grounding phrases such as "yeah", "ok", "mm-hm", "uh-huh", etc. By default, such turns are labeled **neutral**, because groundings serve mainly to encourage another speaker to continue speaking. However:

a) Groundings may occasionally strongly express an emotion (e.g. "yeah!", (sigh) "ok"), thereby overriding the default label.

b) The semantics of certain groundings is associated with weakly expressed understanding, (e.g. "right" and "sure"), and default to *weak positive*.

c) Certain phrases are associated with strongly expressed uncertainty or confusion (e.g. "um" (silence)), and default to **negative**.

Our annotation manual concludes with 8 examples of annotated student turns (as in Figure 1), with links to corresponding audio files. The variety exemplifies how different students express emotions differently at different points in the dialogue, and cover all 6 emotion labels at least once (there are 2 negatives and 2 positives). Also provided is a lengthy audio-enhanced transcript from a single student tutoring dialogue, to exemplify how student emotion changes throughout a single tutoring session. This transcript is shown in part in Figure 2, Appendix A. The transcript is organized in terms of tutor and student turn start and end times. For each student turn, the four Wavesurfer panes are shown.

4.4 Comparison with Prior Schemes

Studies of actor-read speech often make a large number of emotion distinctions, e.g. the LDC Emotional Prosody corpus distinguishes 15 classes. Our work, like other studies of naturally occurring dialogues, uses a more restricted set of emotions, due to the need to first manually annotate such emotions reliably across annotators. As discussed above, our annotation scheme distinguishes negative, neutral, and positive emotions, as well as "weak" and "mixed" classes. Other studies of naturally occurring data have annotated only two emotion classes (e.g. emotional/non-emotional (Batliner et al., 2000), negative/non-negative (Lee et al., 2001)). The study of (Ang et al., 2002) annotates six emotion classes, but collapses most of these for the purposes of emotion prediction.[4] In Section 5, we will similarly explore the impact of collapsing some of our 6 distinctions, to produce simpler 3-way (negative/positive/neutral) and 2-way (negative/non-negative and emotional/non-emotional) schemes.

In further contrast to (Lee et al., 2001), our annotations are context- and task-relative, because like (Ang et al., 2002; Batliner et al., 2003), we are interested in detecting emotional changes across our dialogues. But unlike (Batliner et al., 2003), we allow annotators to be guided by their intuition rather than a set of expected features, to avoid restricting or otherwise influencing their intuitive understanding of emotion expression, and because such features are not used consistently or unambiguously across speakers. Instead, our manual contains annotated audio-enhanced corpus examples (as in Figures 1-2).

5 Analysis of the Annotation Scheme

Given our complete annotation scheme in Section 4, we now explore both the reliability of the scheme at three levels of granularity that have been proposed in prior work, and the accuracy of automatically predicting these variations. These analyses give insight into the tradeoff

[4] (Ang et al., 2002) also discusses the use of an "uncertainty" label, although it did not improve inter-annotator agreement. Our "weak" labels are more similar to an "intensity" dimension found in studies of elicited speech (see (Cowie et al., 2001)).

between interannotator reliability, annotation granularity, and predictive accuracy.

For the purposes of these analyses, we randomly selected 10 transcribed and turn-annotated dialogues from our human-human tutoring corpus (Section 3), yielding 453 student turns from 9 subjects. The turns were separately annotated by two annotators, using the emotion annotation instructions in Section 4. For our machine-learning experiments we follow the methodology in (Litman and Forbes, 2003), instantiated with the learning method (boosted decision trees) and feature set (acoustic-prosodic, lexical, dialogue and contextual) that has given us our best results in ongoing studies.

5.1 Agreed Student Turns

Conflating *Minor* and Neutral Classes

For our first analysis, only our three main emotion classes were distinguished: **negative, neutral, positive**. Our three minor classes, *weak negative, mixed, weak positive*, were conflated with the **neutral** class. A confusion matrix summarizing the resulting inter-annotator agreement is shown in Table 1. The rows correspond to the labels assigned by annotator 1, and the columns correspond to the labels assigned by annotator 2. For example, 90 negatives were agreed upon by both annotators, while 6 negatives assigned by annotator 1 were labeled as neutral by annotator 2. The two annotators agreed on the annotations of 385/453 turns, achieving 84.99% agreement (Kappa = 0.68 (Carletta, 1996)). Such agreement is expected given the difficulty of the task, and exceeds that of prior studies of emotion annotation in naturally occurring speech; (Ang et al., 2002), for example, achieved agreement of 71% (Kappa 0.47), while (Lee et al., 2001) averaged around 70% agreement.

As in (Lee et al., 2001), we next performed a machine learning experiment on the 385 student turns where the two annotators agreed on the emotion label. Our predictive accuracy for this data was 84.75% (using 10 x 10 cross-validation as in (Litman and Forbes, 2003)). Compared to a baseline accuracy of 72.74% achieved by always predicting the majority (neutral) class, our result yields a relative improvement of 44.06%.[5]

	negative	**neutral**	**positive**
negative	90	6	4
neutral	23	280	30
positive	0	5	15

Table 1: Confusion Matrix 1: Minor → Neutral

[5]Relative improvement of x over y = $\frac{error(y)-error(x)}{error(y)}$, where error(x) is 100 - %accuracy(x).

Conflating *Weak* and Negative/Positive Classes

In a second analysis, we again distinguished only our three main emotion classes; however, this time *weak negative* was conflated with **negative**, and *weak positive* was conflated with **positive**. Our *mixed* class was again conflated with **neutral**. A confusion matrix summarizing the resulting inter-annotator agreement is shown in Table 2. As shown, although the number of agreed negative and positive turns increased, overall interannotator agreement decreased to 340/453 turns, or 75.06% (Kappa = 0.60).

We performed our machine learning experiment on these 340 agreed student turns. The predictive accuracy for this data decreased to 79.29%; however, baseline (majority class) accuracy also decreased to 53.24%; thus relative improvement in fact increased to 55.71%

	negative	**neutral**	**positive**
negative	112	9	9
neutral	31	181	53
positive	1	10	47

Table 2: Confusion Matrix 2: Weak → Neg/Pos

Negative/Non-Negative Classes

As Tables 1-2 indicate, our annotators found the **positive** class the most difficult to annotate and agree upon, and the positive class was also the least frequent class overall. Not surprisingly, our prior machine learning experiments have also showed that the positive class is the hardest to predict (Litman and Forbes, 2003). We thus next explored a binary analysis where our positive and neutral classes are conflated, yielding a **negative/non-negative** distinction akin to (Lee et al., 2001). Again however we experimented with conflating our minor weak classes with either the neutral class or their main class counterparts (e.g. weak negative → negative). Two confusion matrices summarizing the resulting inter-annotator agreements are shown in Tables 3 - 4.

In Table 3, our three minor classes are conflated with the neutral class. Interannotator agreement in this case rises sharply to 420/453 turns, or 92.72% (Kappa = 0.80). The predictive accuracy for this data increased to 86.83%; however, baseline (majority class) accuracy also increased to 78.57%; thus relative improvement in fact decreased to 38.54%

	negative	**non-negative**
negative	90	10
non-negative	23	330

Table 3: Confusion Matrix 3: Pos/Neu → Non-Neg

In Table 4, our two weak classes are conflated with their main class counterparts. Interannotator agreement only rises to 403/453 turns, or 88.96% (Kappa = 0.74),

Predictive accuracy decreases to 82.94%. However, baseline (majority class) accuracy also decreases to 72.21%; thus relative improvement was comparable, at 38.61%

	negative	**non-negative**
negative	112	18
non-negative	32	291

Table 4: Conf. Matrix 4: (Weak) Pos/Neu → Non-Neg

Emotional/Non-Emotional Classes

We also explored an alternative binary analysis that conflated our positive and negative classes, yielding an **emotional/non-emotional** distinction, akin to (Batliner et al., 2000). Again we conflated our minor weak classes with either the neutral class or their main class counterparts, as shown in in Tables 5-6. In Table 5, our three minor classes are conflated with the neutral class, yielding agreement on 389/453 turns, or 85.87% (Kappa = 0.67). The predictive accuracy was high at 85.07%, while baseline (majority) accuracy was 71.98%; thus relative improvement was 46.72%

	emotional	**non-emotional**
emotional	109	11
non-emotional	53	280

Table 5: Confusion Matrix 5: Pos/Neg → Emotional

In Table 6, weak classes are conflated with their main class counterparts. Interannotator agreement decreases to 350/453 turns, or 77.26% (Kappa = 0.55). Predictive accuracy was high at 86.14%; moreover, baseline (majority) accuracy was the lowest yet seen, 51.71%, and relative improvement was the best yet seen, at 71.30%

	emotional	**non-emotional**
emotional	169	19
non-emotional	84	181

Table 6: Confusion Matrix 6: (Weak) Pos/Neg → Emo

Summary

A summary of our results across analyses of agreed student turns are shown in Table 7. **NPN** represents analyses distinguishing negative, neutral and positive emotions, **NnN** represents "negative/non-negative" analyses, and **EnE** represents "emotional/non-emotional" analyses. Column "K" shows Kappa for each analysis, "Acc" shows the predictive accuracy achieved by machine learning, "Base" shows the baseline (majority class) accuracy, and "RI" show the relative improvement achieved by learning compared with this baseline.

As can be seen, there is no single optimal way to conflate the original 6 classes; optimality depends on whether maximizing Kappa, predictive accuracy, or expressiveness is most important. For example, conflating minor and neutral labels (the first three rows) yields better annotation reliability than for their counterparts (conflating weak and main labels) in the last three rows; the reverse is true, however, for machine learning performance (measured by relative improvement over the majority class baseline). With respect to expressiveness, only the 3-way NPN distinction can explicitly distinguish positive emotions. With respect to the binary distinctions, annotating negative/non-negative (NnN) can be done most reliably, while predicting emotional/non-emotional (EnE) yields a better relative improvement.

	K	**Acc**	**Base**	**RI**
	minor → neutral			
NPN	.68	84.75%	72.74%	44.06%
NnN	.80	86.83%	78.57%	38.54%
EnE	.67	85.07%	71.98%	46.72%
	weak → main			
NPN	.60	79.29%	53.24%	55.71%
NnN	.74	82.94%	72.21%	38.61%
EnE	.55	86.14%	51.71%	71.30%

Table 7: Summary: Annotation and Learning Results

5.2 Consensus-Labeled Student Turns

Following (Ang et al., 2002), we also explored *consensus labeling*, both to increase our usable data set for prediction, and to include the more difficult annotation cases. For consensus labeling, the original annotators revisited each originally disagreed case, and through discussion, sought a consensus label. Agreement thus rose across all analyses, to 99.12%; we discarded 8/453 turns for lack of consensus. A summary of the consensus labeling across all 6 analyses discussed above is shown in Table 8. The row and column labels are as above, e.g. the **NPN** row represents turns consensus-labeled as negative/neutral/positive, first when all three minor classes are conflated with neutral, and second where the weak minor classes are conflated with their main counterparts.

	minor → neu			**weak → main**		
	neg	**neu**	**pos**	**neg**	**neu**	**pos**
NPN	99	321	25	119	265	61
	neg	**nonneg**		**neg**	**nonneg**	
NnN	99	346		119	326	
	emo	**nonemo**		**emo**	**nonemo**	
EnE	124	321		180	265	

Table 8: Consensus Labeling over Analyses

We performed our machine learning experiment on the consensus data for all 6 analyses. A summary of our

results are shown in Table 9. A comparison of Tables 7-9 shows that for all of our evaluation metrics, our results decrease across all analyses when using consensus data; similar findings were observed in (Ang et al., 2002). While increasing our data set using more difficult examples decreases predictive ability, note that our consensus results are still an improvement over the baseline.

	Acc	Base	RI
	minor → neutral		
NPN	79.97%	72.14%	28.10%
NnN	84.97%	77.75%	32.45%
EnE	80.78%	72.14%	31.01%
	weak → main		
NPN	73.14%	59.55%	33.60%
NnN	81.88%	73.26%	32.24%
EnE	75.75%	59.55%	40.05%

Table 9: Predicting Consensus Labels

6 Extensions to the Analyses

6.1 Minor Emotion Classes

Our analyses so far distinguished only our 3 main emotion classes; our 3 minor classes were always conflated with one or the other of the main classes. In part, this is because our minor labels were consistently employed only later in the development of our scheme; in early versions, annotators optionally labeled the minor classes (in the NOTES pane), for the purpose of post-annotation discussion. At present, only the last 5 of our 10 annotated dialogues are consistently labeled with minor classes. Table 10 shows a confusion matrix for the annotation of all 6 emotion classes for these 5 dialogues. Interannotator agreement is 142/211 turns, or 67.30% (Kappa = 0.54).

Compared to Section 5, we see that this higher level of granularity yields a lower level of agreement. However, most disagreements fall adjacent to the diagonal, indicating that they are mostly differences in strength rather than differences in polarity. The analyses in Section 5 investigated various means of resolving these differences.

	neg	*w. neg*	**neut**	*w. pos*	**pos**	*mix*
neg	**48**	2	0	0	0	2
w. neg	6	**10**	3	2	2	0
neut	2	11	**70**	22	3	3
w. pos	0	1	1	**9**	2	0
pos	0	0	1	1	**1**	0
mix	1	1	2	1	0	**4**

Table 10: Confusion Matrix: All 6 Emotion Classes

6.2 Specific Emotions

Our analyses in Section 5 did not consider the specific emotion annotations in our "EMOb" pane. This is in part because, as with our minor labels, our specific emotion labels were only consistently employed when annotating the last 5 of our 10 dialogues. If we consider only the 66 turns where both annotators agreed that the turn was negative (weak or strong), and view multiple emotion labels which overlap with single emotions as agreed (e.g. sad+bored agrees with a sad or bored label), interannotator agreement is 45/66 turns, or 68.18% (Kappa = 0.41). The same analysis for the 13 positive turns yields 100% agreement (Kappa = 1).

The labels we've included so far are those we've encountered in our human-human tutoring dialogues; we expect to see some differences in the human-computer dialogues, as discussed in Section 6.3, and continue to employ the "other" label. In part, the decision about which specific emotions to ultimately recognize in our system depends on what we want the system to adapt to. This in turn requires some understanding of how human tutors adapt to different emotions. For example, perhaps our tutor responds differently to anger, uncertainty, boredom and confusion, but responds the same to most positive emotions. We are currently investigating this in our annotated human-human tutoring dialogues.

6.3 Human-Computer Corpus

We have just begun annotating our corpus of human-computer spoken tutoring dialogues; to date we have annotated 5 dialogues from 5 different students.

We have applied the 6 reliability analyses in this paper to these annotations, and have found again that most disagreements are simply differences in strength rather than differences in polarity. Our best interannotator reliability was found using the **NnN, weak → main** analysis (contrary to the human-human findings), which gave agreement of 96/115 turns, or 83.48% (Kappa = 0.67).

The corpus example in Figure 4 (Appendix A) highlights differences between our human-human and human-computer tutoring dialogues that potentially might impact emotion annotation. First, both the average student turn length in words, and the average number of student turns per dialogue, are much shorter in the human-computer than in the human-human dialogues. This means that there is less information in the human-computer dialogues to make use of when judging expressed emotions. Second, errors in speech and natural language processing can have a significant effect on the student emotional state in the human-computer tutoring dialogues. Such emotions don't concern either the PHYS domain or the TUT domain, and suggest that we might want to add a third NLP domain if we want the system to respond to these emotions differently. Relatedly, we already see fre-

quency differences across the human-human and human-computer dialogues with respect to specific emotions, for example an increased use of "irritated" in the human-computer data. Finally, computer tutors are far less flexible than human tutors. This alone can effect student emotional state, and furthermore it can limit how the student expresses their own emotional states. For example, in the human-human dialogues we see more student initiative, groundings, and references to prior problems.

7 Conclusions and Current Directions

In this paper we presented and analyzed our scheme for annotating student emotional states in spoken tutoring dialogues. Our scheme distinguishes three main (negative, neutral and positive) and three minor (weak negative, mixed, and weak positive) emotion classes. Our inter-annotator agreement is on par with prior emotion annotation in other types of corpora. We used consensus-labeling to resolve disagreements and increase our dataset. Through further annotation and the use of other inter-annotation metrics (Gwet, 2001), we will investigate how systematic disagreements can yield revisions to our annotation scheme that improve reliability.

Our machine learning experiments have shown that our main emotion categories can be predicted with a high degree of accuracy. Although not presented here, F-Measures ($= \frac{2*Recall*Precision}{Recall+Precision}$) for our experiments on agreed data ranged from 67%-86%; in future work we will more closely examine the tradeoff between recall and precision when predicting our annotations. Our experiments have also highlighted tradeoffs that can be made between coding reliability, predictive accuracy, and annotation scheme granularity.

Finally, we presented initial results in annotating our ITSPOKE human-computer tutoring corpus, and discussed differences from our human-human annotations. This research on emotion annotation and prediction is a first step towards extending the ITSPOKE computer tutoring dialogue system to predict and adapt to student emotional states. Our next goal is to label human tutor reactions to emotional student turns, in order to formulate adaptive strategies for ITSPOKE, and to determine which of our six prediction tasks best triggers adaptation.

Acknowledgments

This research is supported by NSF Grants Nos. 9720359 and No. 0328431. We thank Kurt VanLehn and the Why2-Atlas team, and Scott Silliman of ITSPOKE, for system development 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. 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. of ICSLP*.

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

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), June.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18:32–80, January.

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

K. Gwet. 2001. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company.

H. Holzapfel, C. Fuegen, M. Denecke, and A. Waibel. 2002. Integrating emotional cues into a framework for dialogue management. In *Proc. of ICMI*.

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 *Proc. of ICALT*.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2001. Recognition of negative emotions from the speech signal. In *Proc. of ASRU*.

J. Liscombe, J. Venditti, and J.Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proc. of EuroSpeech*.

D. Litman and K. Forbes. 2003. Recognizing emotion from student speech in tutoring dialogues. In *Proc. of ASRU*.

M. Pantic and L. J. M. Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. of IEEE*, 91(9):1370–1390.

C. P. Rosé and V. Aleven. 2002. Proceedings of the ITS 2002 workshop on empirical methods for tutorial dialogue systems, June.

K. VanLehn, P. W. 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*.

APPENDIX A. Spoken Tutoring Dialogue Corpora Examples

Start	End	Pane	Text in Pane
7.67	8.12	.tutor	ok
196.49	205.80	.student	I dont know about this one at all at first I thought they did have the same amount of force and then I thought they didn't now I don't know
		.EMOa	negative
		.EMOb	confused+sad
		.NOTES	
206.82	211.26	.tutor	um yes at first you thought that force would be same isn't it?
211.60	212.84	.student	yeah I thought so
		.EMOa	negative
		.EMOb	confused+sad
		.NOTES	seems exactly the same as prior turn emotionally
212.19	214.38	.tutor	so then uh why did you change your mind?
214.71	229.49	.student	I don't know I think I thought that they both would at first cause I was thinking that they were both moving in the same direction and then I didn't then I thought well the sun's making the earth move so it has more force
		.EMOa	negative
		.EMOb	confused+sad
		.NOTES	
228.59	258.71	.tutor	no but that is not part of the question it has not been asked which is accelerating more or which is changing in motion or you see in the first question there was a specific question um in the first problem uh previous problem there was a specific question which accelerates more or which which suffers greater change in motion here that has not been asked the only thing asked is about the force whether the force uh earth pulls equally on sun or not that's the only question
258.98	263.69	.student	well I think it does but I don't know why I d- don't I do they move in the same direction I do- don't
		.EMOa	negative
		.EMOb	confused
		.NOTES	
264.13	268.28	.tutor	you see again you see they don't have to move if a force acts on a body
268.47	268.67	.student	it
		.EMOa	weak positive
		.EMOb	enthusiastic
		.NOTES	TUT - interrupts tutor to complete thought
268.77	274.10	.tutor	it does not mean that uh uh I mean it will um
274.41	283.20	.student	if two forces um apply if two forces react on each other then the force is equal it's the Newtons third law
		.EMOa	positive
		.EMOb	confident
		.NOTES	
280.70	290.61	.tutor	um you see the uh actually in this case the motion is there but it is a little complicated motion this is orbital motion
290.77	291.10	.student	mm-hm
		.EMOa	weakly positive
		.EMOb	confident
		.NOTES	
291.69	293.42	.tutor	and uh just as

Figure 2: Annotated Dialogue Excerpt from the Human-Human Spoken Tutoring Dialogue Corpus

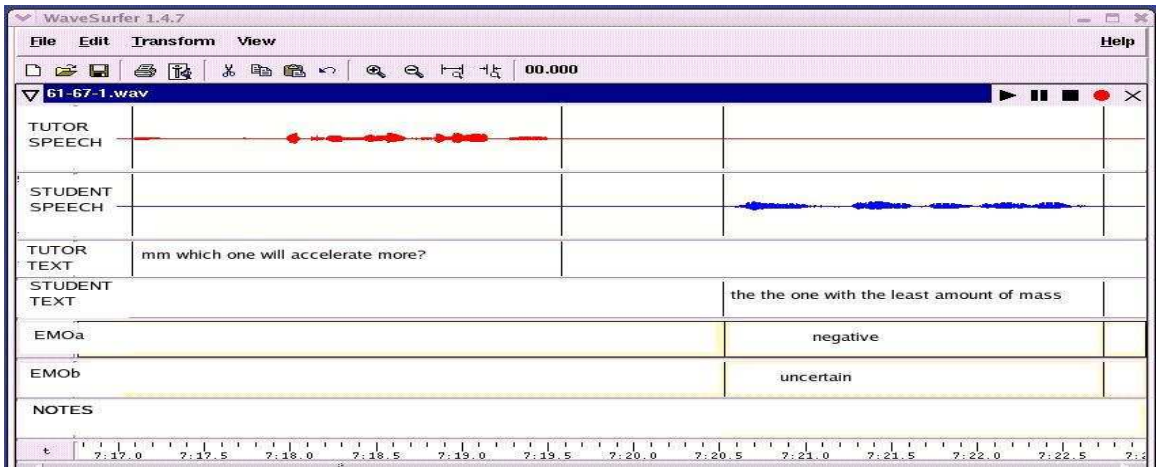

Figure 3: Emotion Annotation using Wavesurfer for the Spoken Tutoring Dialogue Corpora

Start	End	Pane	Text in Pane
301.92	314.65	.tutor	If gravity is the only force acting on an object, will it be moving or staying still?
323.26	326.06	.student	moving
		.EMOa	neutral
		.EMOb	
		.NOTES	
326.25	327.91	.tutor	Could you please repeat that?
328.91	331.55	.student	moving
		.EMOa	neutral
		.EMOb	
		.NOTES	
331.68	335.85	.tutor	I'm sorry, I'm having trouble understanding you. Please try again.
336.98	339.71	.student	moving
		.EMOa	negative
		.EMOb	irritated
		.NOTES	NLP domain
340.62	347.28	.tutor	According to Newton's Second Law, if an object has a non-zero net force acting on it, what will happen to it?
357.58	360.78	.student	it will stay still
		.EMOa	weak negative
		.EMOb	uncertain
		.NOTES	
361.84	371.28	.tutor	Well, to compute the acceleration of a box, one property that you need to know is its mass. What else do you need to know to find the box's acceleration?
376.08	379.06	.student	the direction
		.EMOa	negative
		.EMOb	uncertain
		.NOTES	
379.89	383.51	.tutor	If you see a body accelerate, what caused that acceleration?
386.40	389.30	.student	force
		.EMOa	positive
		.EMOb	confident
		.NOTES	

Figure 4: Annotated Dialogue Excerpt from the ITSPOKE Human-Computer Spoken Tutoring Dialogue Corpus