Using Word-level Features to Better Predict Student Emotions during Spoken Tutoring Dialogues

Abstract

This paper describes our work in developing features and models for detecting student emotional states, given only pitch and lexical information available during a spoken tutoring dialogue. Prior research has primarily focused on the use of turn-level features as predictors. We apply the features at the word level and resolve the problem of combining multiple features per turn using a simple word-level emotion model. Our results show an improvement in prediction using word-level features over using turn-level features. We observe that as turn length increases word-level features constantly outperform turn-level features. Furthermore, our results show that the combination of lexical and pitch features at the word level is a consistent best performer in experiments across corpus-learner combinations.

1 Introduction

We investigate the utility of using prosodic and lexical features applied at the word level for the task of predicting student emotions in two corpora of spoken tutoring dialogues. Motivation for this work comes from the performance gap between human tutors and current machine tutors; typically students tutored by human tutors learn more than students tutored by computer tutors. One of the methods currently being explored as a way of closing this gap is to incorporate affective reasoning into current computer tutoring systems, including dialogue-based tutoring systems, e.g., (Aist et al., 2002; Bhatt et al., 2004; Self, 2004).

Previous spoken dialogue research in other domains has shown that turn-level prosodic, lexical, dialogue, and other features can be used to predict user emotional states (Ang et al., 2002; Devillers et al., 2003; Lee et al., 2001; Shafran et al., 2003). To better approximate the prosodic information (Batliner et al., 2003) uses word-level features and successfully applies them to a different emotion detection task. To our knowledge, there is no previous work that directly compares the impact of using features at the sub-turn rather than the turn level for emotion prediction. In this paper we are performing a first comparison of the two levels for the task of detecting student emotional states.

There are many choices for sub-turn units (breath groups, intonational phrases, syntactic chunks, words, syllables). We will use words as our sub-turn units because it is straightforward to do the segmentation and because these units have been used successfully by other researchers for similar tasks (Batliner et al., 2003). To simplify our word versus turn-level feature comparison, we will focus *only* on lexical and pitch features.

Our hypothesis is that using word-level features will be better for emotion prediction than using turn-level features. The intuition behind this hypothesis is that, at least for pitch information, computing the pitch features at the word level will give a better approximation of the pitch contour, which in turn will help us do better in emotion prediction. Moreover, emotion might not be expressed over an entire turn (especially for long turns) but on certain parts of a student turn; for this reason computing the features at the turn level might mitigate the effect of "emotional" parts of the turn. We will investigate this hypothesis using various corpus-

learner combinations. As our results will show in Section 5, even under a very simple word-level emotion model, using word-level features proves to be as good as and in many cases better than turn level features (especially for pitch feature sets).

Also, to understand how the word-level feature sets affect performance, we will investigate the performance of our turn-level and word-level feature sets as turn length increases. The intuition is that as turn length increases, at least for pitch information, turn-level features give a coarser approximation of the pitch contour (for example: the regression coefficient feature). Our results indicate that the advantage of word-level features lies in a more accurate prediction of longer turns.

We also investigate whether there is a certain feature set that stands out as a robust choice across different corpus-learner combinations. We find that the combination of lexical and pitch features at the word level is a consistent best performer.

2 Emotional speech tutoring corpora

We have developed an annotation scheme for annotating emotions and attitudes in the tutoring domain, and have previously applied it to corpora of both human-human (**HH**) and human-computer (**HC**) tutoring dialogues (Self, 2004). In our annotation scheme, each student turn is labeled for both strong and weak perceived expressions of emotion. *Negative* emotions include emotions like confused, bored, irritated, uncertain and sad, while *positive* emotions include confident and enthusiastic. All other turns are labeled as *neutral*[1].

In our previous work, our three-way annotations and two binary simplifications of it were studied to learn about the ability to predict different types of emotional distinctions: 1) our original classification task (**NPN** – negative / positive / neutral), 2) an Emotional / non-Emotional task (**EnE** – positive and negative are conflated) and 3) a Negative / non-Negative task (**NnN** – positive and neutral are conflated).

In this study we will focus only on the *agreed* turns of the *EnE* annotation scheme. Our ongoing research on the HH corpus suggests that the EnE classification will be useful for triggering system adaptation to student emotions. The agreed turns are the turns labeled with the same emotion class by our two annotators. Following (Self, 2004; Ang et al., 2002), we will use only the agreed subsets of our corpora because they offer the clearest cases of emotional turns.

3 Feature extraction

Conveying the intended meaning of a sentence involves not only appropriate word selection but also the appropriate way of uttering the words. Prosodic features are often computed to quantify this rendering aspect. Recognizing the importance of these two sources of information for the emotion prediction task, we will be using both lexical and prosodic features.

Our lexical features are computed based on a human transcript of the student speech (Table 1). For the turn-level features we used a bag-of-words approach via a word-occurrence vector representation. For the word-level features we used the word itself as the feature. No processing of the transcript (e.g. filtering stop words) was performed.

Pronunciation aspects can be captured using various information sources such as pitch, duration and amplitude. In this paper we will focus only on the pitch information because changes in speaking style are directly reflected in the shape of the pitch contour. Moreover, as we will see in Section 4, word-level pitch features offer a better approximation of the pitch contour shape than turn-level features. If pitch contour shape is indeed useful for emotion prediction, then a better approximation of the contour might result in an improvement in prediction. Since the advantage of word-level features is not that clear for the other prosodic information sources, we elected not to use them in this study. Nonetheless, we believe they are important and we plan to incorporate them in our future work.

Pitch describes how high or low (frequency-wise) speech is rendered. For example, in English, the sentence 'This is great' uttered as an exclamation usually expresses a positive emotion, while the same lexical construct uttered with an alternative pitch contour often expresses a negative emotion.

We will approximate the pitch contour (fundamental frequency or F0) using nine features (see Table 1). Four of them, minimum, maximum, mean and standard deviation, are commonly used

[1] These negative, neutral and positive emotion classes correspond to traditional notions of valence, but these terms are **not** related to the impact of emotion on learning. For example working through negative emotions is hypothesized to be a necessary part of the learning process.

by researchers for various tasks (negative emotion detection– Lee et al.,2001; predicting user corrections– Swerts et al., 2000) and were also employed in previous studies on our corpora (Self, 2004). These four pitch features give us a very coarse approximation of the pitch contour for an entire turn.

Lexical features
• Word occurrence vector (turn-level)
• The word (word-level)
Pitch features
• Minimum
• Maximum
• Mean
• Standard Deviation
• Onset
• Offset
• Linear regression coefficient
• Linear regression error
• Quadratic regression coefficient

Table 1. Features used in our experiments

Inspired by the work of (Batliner et al., 2003), we will use the following new features that offer a better approximation of the pitch contour: onset (the first F0 value), offset (the last F0 value), regression coefficient and regression error. Linear regression is performed to approximate the pitch contour shape. The regression coefficient estimates the direction of the pitch contour (rising or dropping) and can be used to distinguish, for example, questions and statements (at least for English). The regression error offers a better approximation than the standard deviation of the spread of the pitch contour relative to the pitch contour direction (the regression line). Since highly emotional speech is believed to have a large variation in pitch contour, this value may be a good indicator of emotional speech. Moreover, to better approximate the shape of the pitch contour (at least at levels smaller than the entire turn) we also use the second order coefficient of the quadratic interpolation. This value relates to the Tilt model (Taylor, 2000) and approximates the intonation used.

4 Turn and word-level prediction tasks

Recall that our goal is to investigate whether using features at a sub-turn level (word-level in our case) will help in emotion prediction. Using sub-turn features might help in our task due to several reasons. First, computing the pitch features at a smaller level offers a better approximation of the pitch contour than turn-level pitch features (especially regression coefficient and regression error – see Figure 1). Second, emotion might not be expressed during the entire turn (especially for a long turn) but on certain parts of the turn. Returning to our previous 'This is great' example, in general, the word 'great' bears the highest change in prosody between the two styles of rendering the sentence. The small change in prosodic information for the first part of the sentence will mitigate the effect of 'great' if pitch features are computed at the turn level.

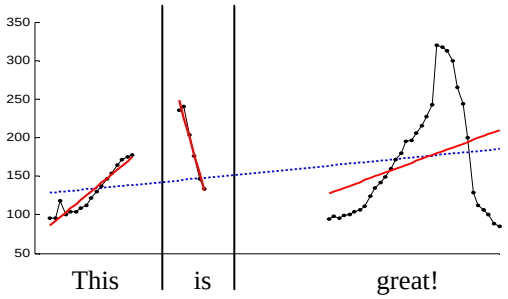

Figure 1. Pitch contour approximation using regression ("This is great" uttered as an exclamation). The regression lines for turn and word level are plotted.

To investigate our hypothesis, we extract the pitch features described in Section 3 at the turn and word level. For the word level features, we first automatically segment our turn-level wave files using the CMU Sphinx 2 speech recognizer running in forced-alignment mode using hand-labeled transcriptions[2]. Then, for each word, instead of using the entire pitch contour, only the segment corresponding to the word in question is used in computing the features. Also, to account for word order and word position in the turn, we create two additional positional features for each word: the number of words before and after that word.

However, using word level features introduces two major problems, given our turn-level annotation scheme. First, we do not know which of the words in an emotional turn are the words where the emotion is expressed. The only thing we know is that the sequence of words results in a certain emotional class. This will impact our training procedure, as discussed below. Second, assuming that we can predict an emotional class for each word, we still need to combine the sequence of predicted

[2] Turns where the automatic word segmentation failed were discarded. No turns were discarded from the HC corpus, but 9% of the HH turns were removed.

emotional classes into a single class (to label the turn as a whole, as in our annotations).

Faced with a similar problem but in the task of speaker identification, (Sönmez et al., 98) use a stylization and regularization algorithm of the pitch contour. Pitch features extracted from sub-turn levels are combined by fitting appropriate parametric distributions. The parameters of these distributions are used as turn/speaker level features in this way bypassing the sub-turn level problems.

In our work, since we are also interested in *incorporating* the lexical information at the sub-turn level, we will use the following simplified *word-level emotion model*. In the training phase, each word is labeled with the turn class and a model for predicting the word emotion is built using all the words from all turns in our training corpus (i.e. we predict word labels). In the test phase, for each turn, we predict the class of each word in the turn and then combine the word classes using majority voting (ties broken randomly). That is, the most frequent emotional class among the turn's words will be the turn's emotional class.

Here is an example from our HC corpus. In the training phase, the student turn "They are the same" will produce *four* training instances, one for each word. Features will be computed for each individual word. In our corpus, this turn was labeled as emotional, thus all four instances will have the emotional class. This training data (which is larger than the training data for turn-level features since many turns have at least two words) is used by the classifier to learn a model. During the test phase, whenever we need to predict the class for a turn, for example the turn "It will change", we will produce an instance for each word in the turn (three instances in our example) and use the learned model to classify them. Finally, the turn class will be the class that labeled the highest number of words in the turn.

5 Results

We will test our hypotheses on four combinations of two contrasting corpora and two contrasting learners. Table 2 highlights some of the differences between our two corpora (the HH and HC corpora). The HC corpus is smaller in size and has shorter turns than the HC one. Conceivably, the HC turns contain less emotional content making prediction more difficult. Our previous turn-level studies (Self, 2004) showed that the two corpora also differ in the types of features that offer the best performance for emotion prediction.

	HH	HC
Number of turns (turn-level instances)	319	220
Number of words (word-level instances)	1310	511
Class distribution (E/nE)	148/171	129/91
Average turn length in words	6.11	2.42
Best accuracy (previous work)	88.86%	66.36%

Table 2. HH and HC corpora properties

As another way to investigate the generality of our results, we use two contrasting learners from the Weka toolkit (Witten and Frank, 1999): a nearest neighbor classifier (IB1) and boosted decision trees (ADA). IB1 is a lazy learner while ADA is an abstraction-based learner. (Rotaru and Litman, 2003) have found that memory-based learning and abstraction-based learning algorithms can produce significantly different performance depending on several factors such as the language learning task, the number of features, and the type of features.

Section 5.1 describes our feature sets. In section 5.2 and 5.3 we report our results using IB1 since it offers the clearest distinction between turn-level and word-level prediction. Section 5.4 discusses our results with ADA.

5.1 Feature sets

Previous work has shown that the addition of lexical information can improve speech based emotion detection (Ang et al., 2002; Lee et al., 2002). Our previous work has shown that, at least for our corpora, the lexical features alone can sometimes outperform the combination. To investigate the knowledge source that yields the highest performance, in this study we create three feature sets for each level: only lexical features, only pitch features and the combination of pitch and lexical features.

At the turn level, the *Lex-Turn* feature set consists of all lexical items in the turn; the turn transcript was converted to a word-occurrence vector representation for this purpose. Next, the *Pitch-Turn* feature set consists of all pitch features described in Section 3 at the turn level. The *PitchLex-Turn* feature set uses both pitch and lexical features at the turn level.

We created corresponding feature sets at the word-level. The *Lex-Word* feature set contains lexical features at the word level. That is, each

word from a turn had as features the word itself plus the two positional features. *Pitch-Word* contains our pitch features at the word level and the two positional features, while *PitchLex-Word* is the combination of both pitch and lexical features at the word level and the two positional features.

As baseline we used the majority class baseline.

5.2 Human-Human corpus, IB1 learner

Figure 2 presents, for each feature set, the mean accuracy and as error bars the confidence intervals[3], computed across 10 runs of 10-fold cross-validation. The bars represent, from left to right, the accuracy using baseline, lexical features at turn and word levels, pitch features at turn and word levels, and the combination of pitch and lexical features at turn and word levels.

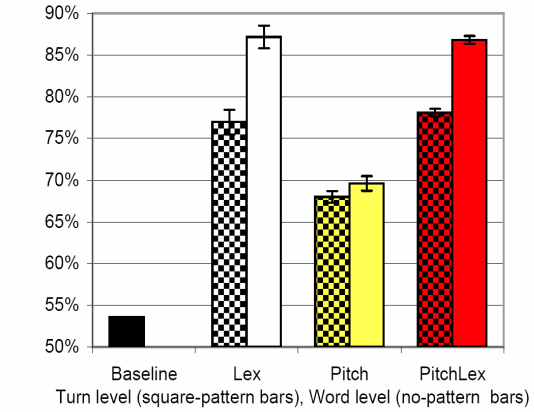

Figure 2. Comparison of turn-level and word-level features (HH corpus, IB1 classifier)

Comparing turn-level feature sets and the equivalent word-level ones, we observe that, at least for the nearest neighbor classifier, using word-level features *always* significantly outperforms turn-level features, which in turn outperforms baseline. If lexical features are present in the feature set, there is a notable increase in performance from turn-level features to word-level features (Lex-Turn vs Lex-Word and PitchLex-Turn vs PitchLex-Word). The improvement on pitch feature sets is of a smaller magnitude but still statistically significant. These observations support our hypothesis that capturing the information at the word-level is helpful for emotion prediction.

[3] Confidence intervals are defined as the mean +/- 2*Standard Error. If the confidence intervals for two bars in the graph intersect then the difference between the two bars is not significant; otherwise the difference is statistically significant with 95% confidence.

We also analyzed the feature sets' performance as a function of turn length to better understand the difference between turn-level and word-level feature sets. Given our small dataset (319 turns), we divided the turns in our corpus in four categories: single (turns with only one word), short (turns with 2 to 4 words), medium (turns with 5 to 10 words) and long (turns with more than 10 words). The distribution in the HH corpus is: single 48%, short 25%, medium 17% and long 10%. Next, we used the predictions from the 10 x 10 cross validation experiments and computed the average accuracy for each of the four categories on all six feature sets (turn-level and word-level)[4].

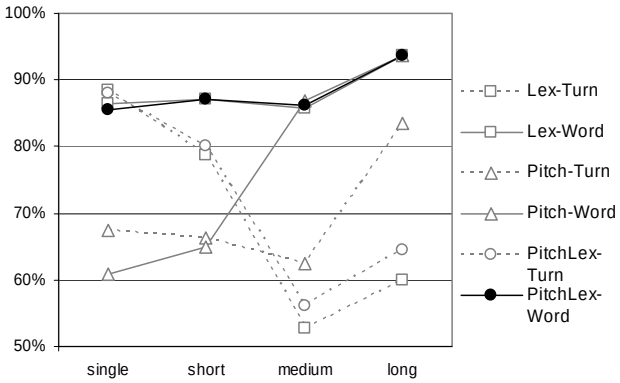

Figure 3. Accuracy as function of turn length (HH corpus, IB1 classifier)

Figure 3 reports the prediction accuracy for each feature set as a function of turn length. For the single category (turn with only one word) we can observe that all turn-level features sets have a better accuracy than the equivalent word-level feature set. For this category, in the test phase all turns have the same feature values for both turn-level and word-level feature sets (the features extracted from that word). We hypothesize that this difference in performance is due to the noise introduced by the word-level emotion model: all words from emotional turns are labeled as emotional. As a result of this, the training set for emotional words is contaminated with words that are actually neutral, resulting in a lower accuracy for word-level features.

We can observe that the lexical feature set at turn-level starts with a relatively high accuracy (88%) and it decreases sharply as turn length in-

[4] Please note that the accuracy values we report for each category are *not* the same with the ones in which we would have trained and tested *only* on turns from that category. Due to corpora size we could not train and test on a single category.

creases; when we reach long turns, the accuracy increases slightly. On the other hand, prosodic feature at turn level start from a lower accuracy (67%) but suffer a much smaller decrease with a big jump in performance for long turns. If we combine the two feature sets (PitchLex-Turn), the contour follows the lexical features' contour with a slight improvement due to the pitch features.

A completely different picture can be observed for word-level features. All word-level features sets exhibit an increasing trend, more notably for pitch features. The difference in performance between Pitch-Turn and Pitch-Word, which becomes more obvious as turn length increases, supports our hypothesis that word-level pitch features give a more accurate account of the pitch information, at least for our emotion prediction task. Interestingly, for medium and long turns all word-level features had almost identical performances.

Going back to Figure 2, consistent with the previous work on the same corpus, if we use turn-level features, we observe that using only pitch features (Pitch-Turn) performs much worse than lexical features (Lex-Turn), but still significantly better than the baseline. Adding the pitch features to the lexical features (PitchLex) improves the performance slightly over lexical features alone, but the difference is not statistically significant. If we use word-level features, the same observations hold: pitch alone does much worse than lexical alone and the combination of the two is comparable with lexical features alone.

We also analyzed which features sets are the best performers. We define best performers as the feature set that yields the highest accuracy plus the feature sets that yield statistically comparable accuracies with the best performer. For the HH corpus, the best performers are Lex-Word and *PitchLex-Word*. The fact that the best performers use only word-level feature sets, at least for the nearest neighbor classifier, supports again our hypothesis that using word-level features is better than using turn-level features.

5.3 Human-Computer corpus, IB1 learner

To explore the generality of our previous results, we reran our experiments on the HC corpus. We used the same emotion annotation (EnE) and the same learner (IB1). The differences between the HC and HH corpora have been highlighted at the beginning of Section 5 and 5.1. Figure 4 summarizes our results on this corpus.

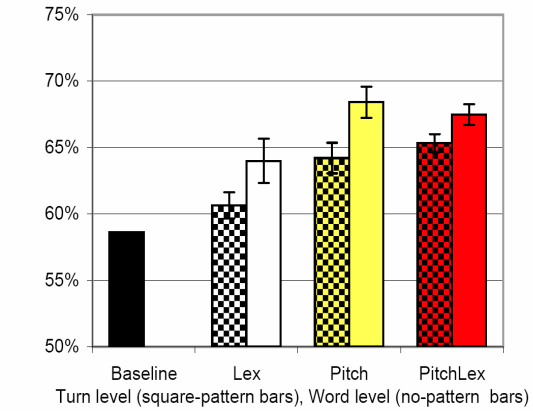

Figure 4. Comparison of turn-level and word-level features (HC corpus, IB1 classifier)

If we compare turn-level feature sets with their word-level counterparts, the results are consistent with the ones on the HH corpus. Whether we use only lexical features, only pitch features or both pitch and lexical features, applying them at the word-level results in a statistical improvement over applying them at the turn-level. This supports again our hypothesis that using word-level features is better than turn-level features.

Figure 5 plots the accuracy for our six feature sets as function of turn length. Given that the turns in this corpus are smaller than in the HC corpus, our categories are: 1 (turns with only one word), 2 (turns with two words), 3 (turns with three words) and more3 (turns with more than three words). The distribution shape is similar to the HH corpus: 1 47%, 2 17%, 3 18% and more3 18%.

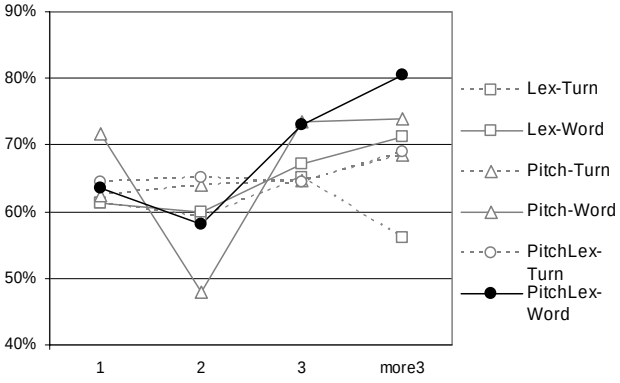

Figure 5. Accuracy as function of turn length (HC corpus, IB1 classifier)

The turn-level performance on this corpus exhibits a somewhat constant trend as turn length increases (slight increase for Pitch-Turn and PitchLex-Turn and a relative decrease for Lex-

Turn). On the other hand, and similar to the observation on the HC corpus, the word-level features exhibit an increasing trend as turn length increases, outperforming all the turn-level feature sets as turn length increases. As in the other corpus, PitchLex-Word places itself as the best performer as turn length increases.

Unlike the HH corpus where the lexical choice was one of the best predictors for emotion, in the HC corpus the lexical feature sets are the lowest performers (although statistically better than the baseline) highlighting again the differences between our corpora.

Our best performers on this corpus are again two word-level feature sets: Pitch-Word and *PitchLex-Word*. Comparing with the best performers from HH corpus, PitchLex-Word is the only feature set that *consistently* generates the best performance or comparable with the best performance. As we will see in Section 5.4 this result generalizes to all corpus-learner combinations.

5.4 Generality of our results

Even though our two corpora had different properties (among them: difference on what type of features helps the most in classification, difference in the turn length), we consistently made the same observations: word-level emotion prediction is better than turn-level prediction, PitchLex-Word is a consistent best performer, and as turn length increases, the advantage of word-level features sets becomes more and more visible. All these observations hold for IB1. In this section we investigate whether our results hold for a contrasting classifier: ADA (boosted decision trees). This classifier consistently yielded the most robust performance across feature sets and corpora in our previous studies (at the turn-level). Table 3 summarizes our results with respect to our hypotheses.

Our results for ADA on the HH corpus are similar to what we observed for IB1: word-level feature sets constantly outperform turn-level features sets. ADA's performance on turn-level features is much better than IB1's performance but the one on word-level features is similar. Even with this boost in performance for turn-level feature sets, word-level feature sets still perform statistically better than their turn-level counterparts (with the exception of the [Lex-Turn, Lex-Word] pair – see the last column of Table 3). If pitch features are part of the feature set pair then the difference is always significant supporting our hypothesis that word-level pitch features help in our emotion prediction task. Again, PitchLex-Word is among the best performers although the list of best performers is larger. Turn length plots show a similar picture: word-level feature sets perform better than turn-level feature sets on longer turns but the difference is not as big as for IB1.

		IB1	ADA
Human-Human	Turn-level VS Word-level		
	Lex	76.96 <* 87.17	86.93 < 88.03
	Pitch	67.98 <* 69.60	74.22 <* 78.67
	PitchLex	78.09 <* 86.81	84.65 <* 87.21
	Best Performers	*PitchLex-Word* Lex-Word	*PitchLex-Word* Lex-Turn Lex-Word
Human-Computer	Turn-level VS Word-level		
	Lex	60.66 <* 64.00	64.85 >* 55.91
	Pitch	64.21 <* 68.41	64.59 < 67.61
	PitchLex	66.87 <* 67.49	66.54 < 66.87
	Best Performers	*PitchLex-Word* Pitch-Word	*PitchLex-Word* Pitch-Turn PitchLex-Turn Pitch-Word

Table 3. Accuracy(%) summary for ADA and IB1 (A "<" sign means that the turn level feature set performs worse than the word level counterpart. A * next to it means the difference is significant)

For the HC corpus, which is a harder task, our results are similar but loose the significance (with the exception of lexical feature sets but here ADA performed worse than baseline for Lex-Word). The loss of significance can be attributed to the fact that ADA's performance on word-level features is actually less than IB1's performance and because ADA's results have bigger variance compared to IB1's ones. Again, PitchLex-Word is one of the best performers.

We would like to mention that the performance on lexical feature sets (Lex-Turn and Lex-Word) is not of great interest for us and we were mostly interested in the performance on pitch feature sets. This is due to the fact that the lexical features at the turn and word level contain almost the same information: they both contain the actual words in the turn. At the turn-level a bag-of-words approach is used; at the word-level there is still a word occurrence approach but it is augmented with the positional features. It is the job of the learner to make sense out of this information and some learners do a better job than others (ADA was much

better than IB1 on turn level lexical features probably due to the fact that it does feature selection). On the other hand, pitch feature information differs significantly between turn and word level as described in the example from Section 4. While the performance of lexical features is dependent on the learner's ability to handle them and the corpus, the performance when pitch features are employed is always consistent: having word-level pitch features is better than turn-level pitch features (in 6 out of 8 cases the difference is statistically significant).

Finally, we would like to mention that the accuracy of our best performers is comparable with the best results reported on our corpora in previous work (Self, 2004). But our work and the previous work can not be compared directly: some differences in the corpus size, we use more pitch features while the previous work uses less pitch features but more prosodic and contextual features.

6 Conclusions and future work

In this paper we have been advocating for the usage of word level lexical and pitch features for emotion prediction in spoken dialogue corpora. We described the problems that we faced when using word level features and addressed them via a word-level emotion model. Even under a very simple word-level emotion model, our results indicate that word-level feature sets perform in general better than the corresponding turn-level feature sets. Our investigation of the performance as function of turn length indicates that word level feature sets handle longer turns much better than the turn level feature sets. We have found that the combination of lexical and pitch features at the word level is a consistent best performer in our corpus-learner combinations. The fact that our results hold for our combinations of contrasting corpora and learners supports the generality of our conclusions.

In our future work we plan to experiment with more refined word-level emotion models. We plan to learn a prosodic model for non-emotional words (based on the assumption that all words from a non-emotional turn are non-emotional) and use it to better identify emotional words in an emotional turn. Then, based on the training set we can learn how to predict the emotional class of a turn from the classes of its words (instead of using majority voting). Filtering irrelevant words (e.g. stop words) might offer further improvements. Finally, integrating other prosodic information sources (amplitude and duration) is among our priorities.

References

G. Aist, B. Kort, R. Reilly, J. Mostow, R. Picard. 2002. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities. In *Proc. of Intelligent Tutoring Systems*.

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

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

K. Bhatt, M. Evens, S. Argamon. 2004. Hedged Responses and Expressions of Affect in Human-Human and Human-Computer Tutorial Interactions. In *Proc. of Cognitive Science.*

L. Devillers, L. Lamel, I. Vasilescu. 2003. Emotion Detection in Task-Oriented Spoken Dialogs. In *Proc. of IEEE Int. Conference on Multimedia & Expo.*

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

Self. 2004.

M. Rotaru, D. Litman. 2003. Exceptionality and Natural Language Learning. In *Proc of Computational Natural Language Learning.*

I. Shafran, M. Riley, M. Mohri. 2003. Voice Signatures. In *Proc. of Automatic Speech Recognition and Understanding.*

K. Sönmez, E. Shriberg, L. Heck, M. Weintraub. 1998. Modeling dynamic prosodic variation for speaker verification. In *Proc. of Int. Conf. Spoken Language Processing.*

M. Swerts, D. Litman, J. Hirschberg. 2000. Corrections in Spoken Dialogue Systems. In *Proc. of Int. Conf. Spoken Language Processing.*

P. Taylor. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America, 107 3:1697—1714*.

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