Co-training for Predicting Emotions with Spoken Dialogue Data

Beatriz Maeireizo and Diane Litman and Rebecca Hwa
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, U.S.A.
beamt@cs.pitt.edu, litman@cs.pitt.edu, hwa@cs.pitt.edu

Abstract

Natural Language Processing applications often require large amounts of annotated training data, which are expensive to obtain. In this paper we investigate the applicability of Co-training to train classifiers that predict emotions in spoken dialogues. In order to do so, we have first applied the wrapper approach with Forward Selection and Naïve Bayes, to reduce the dimensionality of our feature set. Our results show that Co-training can be highly effective when a good set of features are chosen.

1 Introduction

In this paper we investigate the automatic labeling of spoken dialogue data, in order to train a classifier that predicts students' emotional states in a human-human speech-based tutoring corpus. Supervised training of classifiers requires annotated data, which demands costly efforts from human annotators. One approach to minimize this effort is to use Co-training (Blum and Mitchell, 1998), a semi-supervised algorithm in which two learners are iteratively combining their outputs to increase the training set used to re-train each other and generate more labeled data automatically. The main focus of this paper is to explore how Co-training can be applied to annotate spoken dialogues. A major challenge to address is in reducing the dimensionality of the many features available to the learners.

The motivation for our research arises from the need to annotate a human-human speech corpus for the ITSPOKE (Intelligent Tutoring SPOKEn dialogue System) project (Litman and Silliman, 2004). Ongoing research in ITSPOKE aims to recognize emotional states of students in order to build a spoken dialogue tutoring system that automatically predicts and adapts to the student's emotions. ITSPOKE uses supervised learning to predict emotions with spoken dialogue data. Although a large set of dialogues have been collected, only 8% of them have been annotated (10 dialogues with a total of 350 utterances), due to the laborious annotation process. We believe that increasing the size of the training set with more annotated examples will increase the accuracy of the system's predictions. Therefore, we are looking for a less labour-intensive approach to data annotation.

2 Data

Our data consists of the student turns in a set of 10 spoken dialogues randomly selected from a corpus of 128 qualitative physics tutoring dialogues between a human tutor and University of Pittsburgh undergraduates. Prior to our study, the 453 student turns in these 10 dialogues were manually labeled by two annotators as either "Emotional" or "Non-Emotional" (Litman and Forbes-Riley, 2004). Perceived student emotions (e.g. confidence, confusion, boredom, irritation, etc.) were coded based on both what the student said and how he or she said it. For this study, we use only the 350 turns where both annotators agreed on the emotion label. 51.71% of these turns were labeled as non-emotional and the rest as emotional.

Also prior to our study, each annotated turn was represented as a vector of 449 features hypothesized to be relevant for emotion prediction (Forbes-Riley and Litman, 2004). The features represent acoustic-prosodic (pitch, amplitude, temporal), lexical, and other linguistic characteristics of both the turn and its local and global dialogue context.

3 Machine Learning Techniques

In this section, we will briefly describe the machine learning techniques used by our system.

3.1 Co-training

Blum and Mitchell (1998) proposed Co-training as a novel method that addresses the challenge of boosting the performance of a learning algorithm using a large set of unlabeled data when we only have a small set of labeled examples. Co-training uses two learning algorithms that train independently on distinct views of the examples to

classify the unlabeled data, in order to enlarge the training set of the other.

In this research, we have developed a learner for each of the two classes: Emotional and Non Emotional. Each of the learners aims to label the examples of its class as precisely as possible.

The algorithm for our Co-training System is shown in figure 1. Each learner selects the examples whose predicted labeled corresponds to its expertise class with the highest confidence. The maximum number of iterations and the number of examples added per iteration are parameters of the system.

```
While iteration < MAXITERATION
   Emo_Learner.Train(train)
   NE_Learner.Train(train)

   emo_Predictions = Emo_Learner.Predict(predict)
   ne_Predictions = NE_Learner.Predict(predict)

   emo_sorted_Predictions = Sort_by_confidence(
                            emo_Predictions)
   ne_sorted_Predictions = Sort_by_confidence(
                            ne_Predictions)

   best_emo = Emo_Learner.select_best(
                            emo_sorted_Predictions,
                            NUM_SAMPLES_TO_ADD)
   best_ne = NE_Learner.select_best(
                            ne_sorted_Predictions,
                            NUM_SAMPLES_TO_ADD)

   train = train ∪ best_emo ∪ best_ne
   predict = predict - best_emo - best_ne
end
```

Figure 1. Algorithm for Co-training System

3.2 Wrapper Approach with Forward Selection

As described in Section 2, 449 features have been currently extracted from each utterance of the ITSPOKE corpus (where an utterance is a student's turn in a dialogue). Unfortunately, high dimensionality, i.e. large amount of input features, may lead to a large variance of estimates, noise, overfitting, and in general, higher complexity and inefficiencies in the learners. Different approaches have been proposed to address this problem. In this work, we have used the Wrapper Approach with Forward Selection.

The Wrapper Approach, introduced by John et al. (1994) and refined later by Kohavi and John (1997), is a method that searches for a good sub-set of relevant features using an induction algorithm as part of the evaluation function. We can apply different search algorithms to find this set of features.

Forward Selection is a greedy search algorithm that begins with an empty set of features, and greedily adds features to the set. Figure 2 shows our algorithm implemented for the forward wrapper approach

```
bestFeatures = []
while dim(bestFeatures) < MINFEATURES
  for iterations = 1: MAXITERATIONS
        split train into training/development
        parameters = computeParameters(training)
        for feature = 1:MAXFEATURES

             evaluate(parameters,development,
                      [bestFeatures + feature])

             keep validation performance
        end

      end
      average_performance and keep average_performance
   end
   B = best average_performance
   bestFeatures ← B ∪ bestFeatures
end
```

Figure 2. Implemented algorithm for forward wrapper approach. The variables underlined are the ones whose parameters we have changed in order to test and improve the performance.

We can use different criteria to select the feature to add, depending on the object of optimization.

The first criterion we are interested in is the accuracy in predicting the student emotional state from the training set generated by the Co-training system. Thus, we need to find the best set of features for accuracy.

Earlier, we have explained the basis of the Co-training system. When developing an expert learner in one class, we want it to be correct most of the time when it guesses that class. That is, we want the classifier to have high precision (possibly at the cost of lower overall accuracy). Therefore, we are interested in finding the best set of features for precision in each class. In this case, we are focusing on emotional and non-emotional classifiers.

Figure 3 shows the formulas used for the optimization criterion on each class. For the Emotional Class, our optimization criterion was to maximize the PPV (Positive Predictive Value), and for the Non-Emotional Class our optimization criterion was to maximize the NPV (Negative Predictive Value).

		MODEL	
		1	0
TARGET	1	TP	FN
	0	FP	TN

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Figure 3. Confusion Matrix, Positive Predictive Value (Precision for Emotional) and Negative Predictive Value (Precision for Non-Emotional)

4 Experiments

For the following experiments, we fixed the size of our test set to 140 examples (40%), and the size of our training set to 175 examples (50%). The remaining 10% has been saved for later experiments.

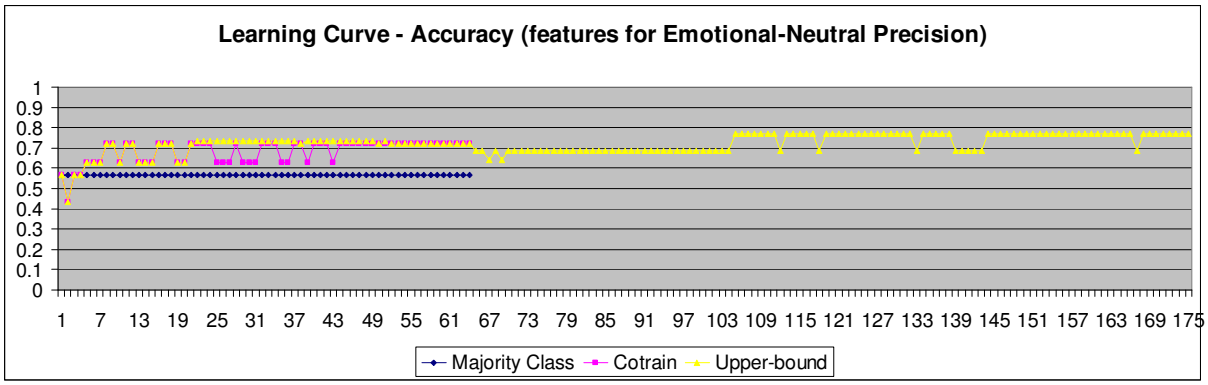

Figure 4. Learning Curve of Accuracy using best features for Precision of Emotional/Non-Emotional

4.1 Selecting the features

The first task was to reduce the dimensionality and find the best set of features for maximizing the accuracy, PPV for emotional class and NPV for non-emotional class. We applied the Wrapper Approach with Forward Selection as described in section 2.2, using Naïve Bayes to get the evaluation measurements for each subset of features.

We have used 175 examples for the training set (used to select the best features) and 140 for the test set (used to measure the performance). The training set is randomly divided into two sets in each iteration of the algorithm: One for training and the other for development (65% and 35% respectively). We train the learners with the training set and we evaluate the performance to pick the best feature with the development set.

Number of Features	Naïve Bayes	AdaBoost-j48 Decision Trees
All Features	74.29 %	87.14 %
10	87.86 %	97.86 %

Table 1. Accuracy with all features and 10 best features using Naïve Bayes and AdaBoost-j48 Decision Trees

The selected features that gave the best accuracy are 60% lexical items, 20% acoustic-prosodic features and 20% other acoustic features. By using them, we increased the accuracy of Naïve Bayes from 74.29% (using all features) to 87.86%, and of AdaBoost-j48 Decision Trees from 87.14% to 97.86%. (See table 1).

By selecting features to optimize PPV for Emotional Class, we increased the precision from 77.19% to 90.95% using 2 lexical features and one acoustic-prosodic feature.

For the Non-Emotional Class, we achieved 100% precision just by using one lexical feature, and it remained the same with the set of 3 best features, one lexical and two non-acoustic prosodic features.

These two set of features for each learner are disjoint.

4.2 Co-training experiments

The two learners are initialized with only 6 labeled examples in the training set, 140 "pseudo-labeled" examples[1] in the Prediction Set and 140 instances in the Test Set. The size of the training set increased each iteration, by adding the 2 best examples (those with the highest confidence scores) labeled by the two learners. The emotional learner and the Non-emotional learner were set to work with the set of features selected by the wrapper approach as described in section 3.1.

We have used Weka's (Witten and Frank, 2000) AdaBoost's version of j48 decision trees (as used in Forbes-Riley and Litman, 2004) to generate the learning curves presented next.

Figure 4 illustrates the learning curve for the accuracy, taking the features selected to label the examples. We used the 3 best features for PPV (see Section 4.1) for the Emotional Learner and the best feature for NPV for the Non-Emotional Learner (see Section 4.1). The x-axis shows the number of training examples added; the y-axis shows the accuracy of the classifier on test instances. We compare the learning curve from Co-training with a baseline of majority class and an upper-bound, in which the classifiers are trained on human-annotated data. Post-hoc analyses reveal that four incorrectly labeled examples were added to the training set: example numbers 21, 22, 45, and 51 (see the x-axis). Shortly after the inclusion of example 21, the Co-training learning curve diverges from the upper-bound. All of them correspond to non-emotional examples that were labeled as emotional by the emotional learner with the highest confidence.

[1] This means that although the example has been labeled, the label remains unseen to the learners.

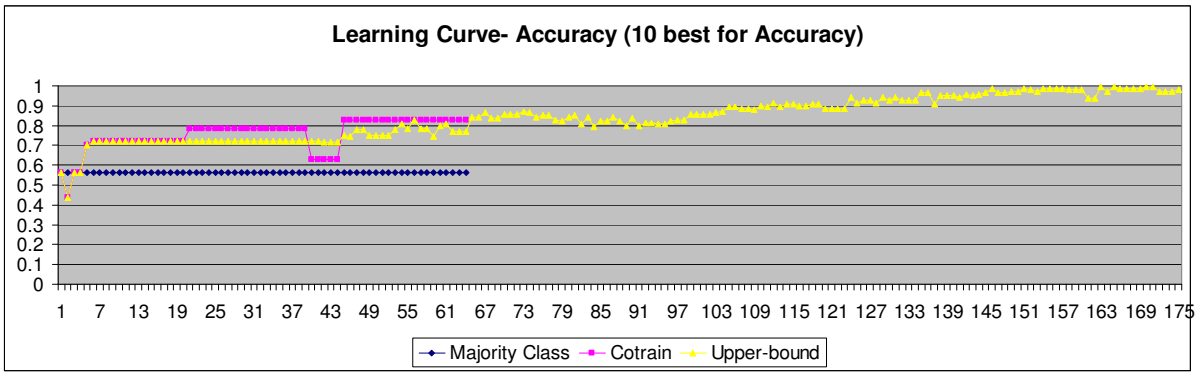

Figure 5. Learning Curve of Accuracy using 10 best features for Accuracy

The Co-training system stopped after adding 58 examples to the training set due to the noise of the results, i.e. the learners were not able to label the examples of their expertise. However, as we can see, the training set generated by the Co-training technique can perform almost as well as the upper-bound, even if incorrectly labeled examples are included in the training set.

The results were even more encouraging when applying the 10 best features for accuracy (see in Section 4.1) to compare the performance of the examples labelled by the Co-training system versus the manually labeled examples. As figure 5 shows, the training set generated by the Co-training system performed even better than the manually labeled training-set. Again, the discrepancy between the results with manually labeled data and the Co-training generated data appeared when noisy incorrect examples were added to the training set.

5 Conclusion

We have shown Co-training to be applicable for predicting emotions from spoken dialogue data. We have given an algorithm that has achieved a high level of accuracy with very few manually labeled examples. Although some noise was introduced to the training set when examples were automatically labeled, the distortion did not significantly degrade the accuracy of the predictions.

We have shown the positive effect of selecting a good set of features optimizing precision and accuracy, and we have shown that the features can be identified with the Wrapper Approach.

In the future, we will try to address the limitation of noise in the learners of the Co-training System. We will also try to generalize our solution to a corresponding corpus of human-computer data (Litman and Forbes-Riley, 2004), and will conduct experiments comparing Co-training with other semi-supervised approaches such as self-training and active learning.

6 Acknowledgements

Thanks to Richard Pelikan, Tomas Singliar and Milos Hauskrecht for their contribution with Feature Selection. This research is partially supported by NSF Grant No. 0328431.

References

A. Blum and T. Mitchell. 1998. *Combining Labeled and Unlabeled Data with Co-training.* Proceedings of the 11th Annual Conference on Computational Learning Theory: 92-100.

A. Blum and P. Langley. 1997. *Selection of Relevant Features and Examples in Machine Learning.* Artificial Intelligence: 245-272.

K. Forbes-Riley and D. Litman. 2004. *Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources.* Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).

G. H. John, R. Kohavi and K. Pleger 1994. *Irrelevant Features and the Subset Selection Problem.* Machine Learning: Proceedings of 11th International Conference:121-129, Morgan Kaufmann Publishers, San Francisco, CA.

R. Kohavi and G. H. John. 1997. *Wrappers for Feature Subset Selection.* Artificial Intelligence, Volume 97, Issue 1-2.

D. J. Litman and K. Forbes-Riley, 2004. *Annotating Student Emotional States in Spoken Tutoring Dialogues.* Proc. 5th Special Interest Group on Discourse and Dialogue Workshop on Discourse and Dialogue (SIGdial).

D. J. Litman and S. Silliman, 2004. *ITSPOKE: An Intelligent Tutoring Spoken Dialogue System.* Companion Proceedings of Human Language Technology conf. of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)