Semi-supervised Learning for Automatic Prosodic Event Detection Using Co-training Algorithm

Je Hun Jeon and Yang Liu
Computer Science Department
The University of Texas at Dallas, Richardson, TX, USA
{jhjeon,yangl}@hlt.utdallas.edu

Abstract

Most of previous approaches to automatic prosodic event detection are based on supervised learning, relying on the availability of a corpus that is annotated with the prosodic labels of interest in order to train the classification models. However, creating such resources is an expensive and time-consuming task. In this paper, we exploit semi-supervised learning with the co-training algorithm for automatic detection of coarse level representation of prosodic events such as pitch accents, intonational phrase boundaries, and break indices. We propose a confidence-based method to assign labels to unlabeled data and demonstrate improved results using this method compared to the widely used agreement-based method. In addition, we examine various informative sample selection methods. In our experiments on the Boston University radio news corpus, using only a small amount of the labeled data as the initial training set, our proposed labeling method combined with most confidence sample selection can effectively use unlabeled data to improve performance and finally reach performance closer to that of the supervised method using all the training data.

1 Introduction

Prosody represents suprasegmental information in speech since it normally extends over more than one phoneme segment. Prosodic phenomena manifest themselves in speech in different ways, including changes in relative intensity to emphasize specific words or syllables, variations of the fundamental frequency range and contour, and subtle timing variations, such as syllable lengthening and insertion of pause. In spoken utterances, speakers use prosody to convey emphasis, intent, attitude, and emotion. These are important cues to aid the listener for interpretation of speech. Prosody also plays an important role in automatic spoken language processing tasks, such as speech act detection and natural speech synthesis, because it includes aspect of higher level information that is not completely revealed by segmental acoustics or lexical information.

To represent prosodic events for the categorical annotation schemes, one of the most popular labeling schemes is the Tones and Break Indices (ToBI) framework (Silverman et al., 1992). The most important prosodic phenomena captured within this framework include pitch accents (or prominence) and prosodic phrase boundaries. Within the ToBI framework, prosodic phrasing refers to the perceived grouping of words in an utterance, and accent refers to the greater perceived strength or emphasis of some syllables in a phrase. Corpora annotated with prosody information can be used for speech analysis and to learn the relationship between prosodic events and lexical, syntactic and semantic structure of the utterance. However, it is very expensive and time-consuming to perform prosody labeling manually. Therefore, automatic labeling of prosodic events is an attractive alternative that has received attention over the past decades. In addition, automatically detecting prosodic events also benefits many other speech understanding tasks.

Many previous efforts on prosodic event detection were supervised learning approaches that used acoustic, lexical, and syntactic cues. However, the major drawback with these methods is that they require a hand-labeled training corpus and depend on specific corpus used for training. Limited research has been conducted using unsupervised and semi-supervised methods. In this paper, we exploit semi-supervised learning with the

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 540–548,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

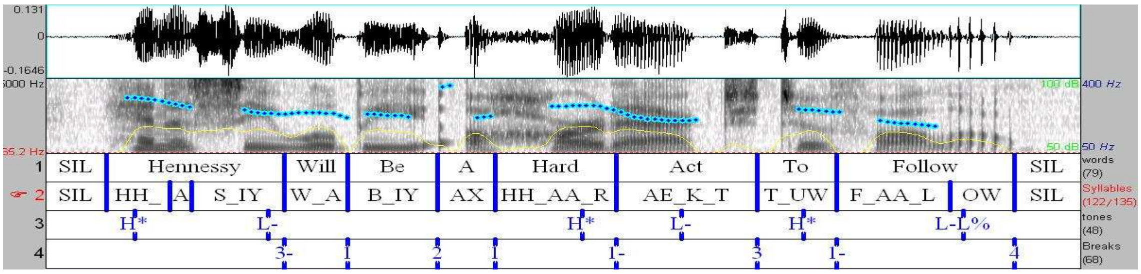

Figure 1: An example of ToBI annotation on a sentence *"Hennessy will be a hard act to follow."*

co-training algorithm (Blum and Mitchell, 1998) for automatic prosodic event labeling. Two different views according to acoustic and lexical-syntactic knowledge sources are used in the co-training framework. We propose a confidence-based method to assign labels to unlabeled data in training iterations and evaluate its performance combined with different informative sample selection methods. Our experiments on the Boston Radio News corpus show that the use of unlabeled data can lead to significant improvement of prosodic event detection compared to using the original small training set, and that the semi-supervised learning result is comparable with supervised learning with similar amount of training data.

The remainder of this paper is organized as follows. In the next section, we provide details of the corpus and the prosodic event detection tasks. Section 3 reviews previous work briefly. In Section 4, we describe the classification method for prosodic event detection, including the acoustic and syntactic prosodic models, and the features used. Section 5 introduces the co-training algorithm we used. Section 6 presents our experiments and results. The final section gives a brief summary along with future directions.

2 Corpus and tasks

In this paper, our experiments were carried out on the Boston University Radio News Corpus (BU) (Ostendorf et al., 2003) which consists of broadcast news style read speech and has ToBI-style prosodic annotations for a part of the data. The corpus is annotated with orthographic transcription, automatically generated and hand-corrected part-of-speech (POS) tags, and automatic phone alignments.

The main prosodic events that we are concerned to detect automatically in this paper are phrasing and accent (or prominence). Prosodic phrasing refers to the perceived grouping of words in an utterance, and prominence refers to the greater perceived strength or emphasis of some syllables in a phrase. In the ToBI framework, the pitch accent tones (*) are marked at every accented syllable and have five types according to pitch contour: H*, L*, L*+H, L+H*, H+!H*. The phrase boundary tones are marked at every intermediate phrase boundary (L-, H-) or intonational phrase boundary (L-L%, L-H%, H-H%, H-L%) at certain word boundaries. There are also the break indices at every word boundary which range in value from 0 through 4, where 4 means intonational phrase boundary, 3 means intermediate phrase boundary, and a value under 3 means phrase-medial word boundary. Figure 1 shows a ToBI annotation example for a sentence *"Hennessy will be a hard act to follow."* The first and second tiers show the orthographic information such as words and syllables of the utterance. The third tier shows the accents and phrase boundary tones. The accent tone is located on each accented syllable, such as the first syllable of word *"Hennessy."* The boundary tone is marked on every final syllable if there is a prosodic boundary. For example, there are intermediate phrase boundaries after words *"Hennessy"* and *"act"*, and there is an intonational phrase boundary after word *"follow."* The fourth tier shows the break indices at the end of every word.

The detailed representation of prosodic events in the ToBI framework creates a serious sparse data problem for automatic prosody detection. This problem can be alleviated by grouping ToBI labels into coarse categories, such as presence or absence of pitch accents and phrasal tones. This also significantly reduces ambiguity of the task. In this paper, we thus use coarse representation (presence versus absence) for three prosodic event detection tasks:

- Pitch accents: accent mark (*) means presence.

- Intonational phrase boundaries (IPB): all of the IPB tones (%) are grouped into one category.

- Break indices: value 3 and 4 are grouped together to represent that there is a break. This task is equivalent to detecting the presence of intermediate and intonational phrase boundaries.

These three tasks are binary classification problems. Similar setup has also been used in other previous work.

3 Previous work

Many previous efforts on prosodic event detection used supervised learning approaches. In the work by Wightman and Ostendorf (1994), binary accent, IPB, and break index were assigned to syllables based on posterior probabilities computed from acoustic evidence using decision trees, combined with a bigram model of accent and boundary patterns. Their method achieved an accuracy of 84% for accent, 71% for IPB, and 84% for break index detection at the syllable level. Chen et al. (2004) used a Gaussian mixture model for acoustic-prosodic information and neural network based syntactic-prosodic model and achieved pitch accent detection accuracy of 84% and IPB detection accuracy of 90% at the word level. The experiments of Ananthakrishnan and Narayanan (2008) with neural network based acoustic-prosodic model and a factored n-gram syntactic model reported 87% accuracy on accent and break index detection at the syllable level. The work of Sridhar et al. (2008) using a maximum entropy model achieved accent and IPB detection accuracies of 86% and 93% on the word level.

Limited research has been done in prosodic detection using unsupervised or semi-supervised methods. Ananthakrishnan and Narayanan (2006) proposed an unsupervised algorithm for prosodic event detection. This algorithm was based on clustering techniques to make use of acoustic and syntactic cues and achieved accent and IPB detection accuracies of 77.8% and 88.5%, compared with the accuracies of 86.5% and 91.6% with supervised methods. Similarly, Levow (2006) tried clustering based unsupervised approach on accent detection with only acoustic evidence and reported accuracy of 78.4% for accent detection compared with 80.1% using supervised learning. She also exploited a semi-supervised approach using Laplacian SVM classification on a small set of examples. This approach achieved 81.5%, compared to 84% accuracy for accent detection in a fully supervised fashion.

Since Blum and Mitchell (1998) proposed co-training, it has received a lot of attention in the research community. This multi-view setting applies well to learning problems that have a natural way to divide their features into subsets, each of which are sufficient to learn the target concept. Theoretical and empirical analysis has been performed for the effectiveness of co-training such as Blum and Mitchell (1998), Goldman and Zhou (2000), Nigam and Ghani (2000), and Dasuta et al. (2001). More recently, researchers have begun to explore ways of combing ideas from sample selection with that of co-training. Steedman et al. (2003) applied co-training method to statistical parsing and introduced sample selection heuristics. Clark et al. (2003) and Wang et al. (2007) applied co-training method in POS tagging using agreement-based selection strategy. Co-testing (Muslea et al., 2000), one of active learning approaches, has a similar spirit. Like co-training, it consists of two classifiers with redundant views and compares their outputs for an unlabeled example. If they disagree, then the example is considered as a contention point, and therefore a good candidate for human labeling.

In this paper, we apply co-training algorithm to automatic prosodic event detection and propose methods to better select samples to improve semi-supervised learning performance for this task.

4 Prosodic event detection method

We model the prosody detection problem as a classification task. We separately develop acoustic-prosodic and syntactic-prosodic models according to information sources and then combine the two models. Our previous supervised learning approach (Jeon and Liu, 2009) showed that a combined model using Neural Network (NN) classifier for acoustic-prosodic evidence and Support Vector Machine (SVM) classifier for syntactic-prosodic evidence performed better than other classifiers. We therefore use NN and SVM in this study. Note that our feature extraction is performed at the syllable level. This is straightforward for accent detection since stress is defined associated with syllables. In the case of IPB and break index detection, we use only the features from the final syllable of a word since those events are associated with word boundaries.

4.1 The acoustic-prosodic model

The most likely sequence of prosodic events $P^* = \{p_1^*, \ldots, p_n^*\}$ given the sequence of acoustic evidences $A = \{a_1, \ldots, a_n\}$ can be found as following:

$$\begin{aligned} P^* &= \arg\max_P p(P|A) \\ &\approx \arg\max_P \prod_{i=1}^{n} p(p_i|a_i) \end{aligned} \quad (1)$$

where $a_i = \{a_i^1, \ldots, a_i^t\}$ is the acoustic feature vector corresponding to a syllable. Note that this assumes that the prosodic events are independent and they are only dependent on the acoustic observations in the corresponding locations.

The primary acoustic cues for prosodic events are pitch, energy and duration. In order to reduce the effect by both inter-speaker and intra-speaker variation, both pitch and energy values were normalized (z-value) with utterance specific means and variances. The acoustic features used in our experiments are listed below. Again, all of the features are computed for a syllable.

- Pitch range (4 features): maximum pitch, minimum pitch, mean pitch, and pitch range (difference between maximum and minimum pitch).

- Pitch slope (5 features): first pitch slope, last pitch slope, maximum plus pitch slope, maximum minus pitch slope, and the number of changes in the pitch slope patterns.

- Energy range (4 features): maximum energy, minimum energy, mean energy, and energy range (difference between maximum and minimum energy).

- Duration (3 features): normalized vowel duration, pause duration after the word final syllable, and the ratio of vowel durations between this syllable and the next syllable.

Among the duration features, the pause duration and the ratio of vowel durations are only used to detect IPB and break index, not for accent detection.

4.2 The syntactic-prosodic model

The prosodic events P^* given the sequence of lexical and syntactic evidences $S = \{s_1, \ldots, s_n\}$ can be found as following:

$$\begin{aligned} P^* &= \arg\max_P p(P|S) \\ &\approx \arg\max_P \prod_{i=1}^{n} p(p_i|\phi(s_i)) \end{aligned} \quad (2)$$

where $\phi(s_i)$ is chosen such that it contains lexical and syntactic evidence from a fixed window of syllables surrounding location i.

There is a very strong correlation between the prosodic events in an utterance and its lexical and syntactic structure. Previous studies have shown that for pitch accent detection, the lexical features such as the canonical stress patterns from the pronunciation dictionary perform better than the syntactic features, while for IPB and break index detection, the syntactic features such as POS work better than the lexical features. We use different feature types for each task and the detailed features are as follows:

- Accent detection: syllable identity, lexical stress (exist or not), word boundary information (boundary or not), and POS tag. We also include syllable identity, lexical stress, and word boundary features from the previous and next context window.

- IPB and Break index detection: POS tag, the ratio of syntactic phrases the word initiates, and the ratio of syntactic phrases the word terminates. All of these features from the previous and next context windows are also included.

4.3 The combined model

The two models above can be coupled as a classifier for prosodic event detection. If we assume that the acoustic observations are conditionally independent of the syntactic features given the prosody labels, the task of prosodic detection is to find the optimal sequence P^* as follows:

$$P^* = \arg\max_P p(P|A, S)$$

$$\approx \arg\max_{P} p(P|A)p(P|S)$$

$$\approx \arg\max_{P} \prod_{i=1}^{n} p(p_i|a_i)^{\lambda} p(p_i|\phi(s_i)) \quad (3)$$

where λ is a parameter that can be used to adjust the weighting between syntactic and the acoustic model. In our experiments, the value of λ is estimated based on development data.

5 Co-training strategy for prosodic event detection

Co-training (Blum and Mitchell, 1998) is a semi-supervised multi-view algorithm that uses the initial training set to learn a (weak) classifier in each view. Then each classifier is applied to all the unlabeled examples. Those examples that each classifier makes the most confident predictions are selected and labeled with the estimated class labels and added to the training set. Based on the new training set, a new classifier is learned in each view, and the whole process is repeated for some iterations. At the end, a final hypothesis is created by combining the predictions of the classifiers learned in each view.

As described in Section 4, we use two classifiers for the prosodic event detection task based on two different information sources: one is the acoustic evidence extracted from the speech signal of an utterance; the other is the lexical and syntactic evidence such as syllables, words, POS tags and phrasal boundary information. These are two different views for prosodic event detection and fit the co-training framework.

The general co-training algorithm we used is described in Algorithm 1. Given a set *L* of labeled data and a set *U* of unlabeled data, the algorithm first creates a smaller pool *U′* containing *u* unlabeled data. It then iterates in the following procedure. First, we use *L* to train two distinct classifiers: the acoustic-prosodic classifier *h1*, and the syntactic classifier *h2*. These two classifiers are used to examine the unlabeled set *U′* and assign *"possible"* labels. Then we select some samples to add to *L*. Finally, the pool *U′* is recreated from *U* at random. This iteration continues until reaching the defined number of iterations or *U* is empty.

The main issue of co-training is to select training samples for next iteration so as to minimize noise and maximize training utility. There are two issues: (1) the accurate self-labeling method for unlabeled data and (2) effective heuristics to select more informative examples. We investigate different approaches to address these issues for the prosodic event detection task. The first issue is how to assign possible labels accurately. The general method is to let the two classifiers predict the class for a given sample, and if they agree, the hypothesized label is used. However, when this agreement-based approach is used for prosodic event detection, we notice that there is not only difference in the labeling accuracy between positive and negative samples, but also an imbalance of the self-labeled positive and negative examples (details in Section 6). Therefore we believe that using the hard decisions from the two classifiers along with the agreement-based rule is not enough to label the unlabeled samples. To address this problem, we propose an approximated confidence measure based on the combined classifier (Equation 3). First, we take a squared root of the classifier's posterior probabilities for the two classes, denoted as $score(pos)$ and $score(neg)$, respectively. Our proposed confidence is the distance between these two scores. For example, if the classifier's hypothesized label is positive, then:

Positive confidence=score(pos)-score(neg)

Similarly if the classifier's hypothesis is negative, we calculate a negative confidence:

Negative confidence=score(neg)-score(pos)

Then we apply different thresholds of confidence level for positive and negative labeling. The thresholds are chosen based on the accuracy distribution obtained on the labeled development data and are reestimated at every iteration. Figure 2 shows the accuracy distribution for accent detection according to different confidence levels in the first iteration. In Figure 2, if we choose 70% labeling accuracy, the positive confidence level is about

Algorithm 1 General co-training algorithm.

```
Given a set L of labeled training data and a set U of unlabeled data
Randomly select U′ from U, |U′|=u
while iteration < k do
  Use L to train classifiers h1 and h2
  Apply h1 and h2 to assign labels for all examples in U′
  Select n self-labeled samples and add to L
  Remove these n samples from U
  Recreate U′ by choosing u instances randomly from U
end while
```

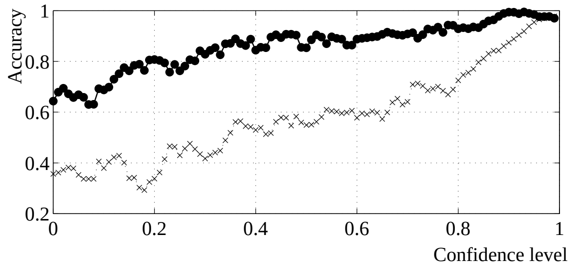

Figure 2: Approximated confidence level and labeling accuracy on accent detection task.

0.1 and the negative confidence level is about 0.8. In our confidence-based approach, the samples with a confidence level higher than these thresholds are assigned with the classifier's hypothesized labels, and the other samples are disregarded.

The second problem in co-training is how to select informative samples. Active learning approaches, such as Muslea et al. (2000), can generally select more informative samples, for example, samples for which two classifiers disagree (since one of two classifiers is wrong) and ask for human labels. Co-training approaches cannot, however, use this selection method since there is a risk to label the disagreed samples. Usually co-training selects samples for which two classifiers have the same prediction but high difference in their confidence measures. Based on this idea, we applied three sampling strategies on top of our confidence-based labeling method:

- Random selection: randomly select samples from those that the two classifiers have different posterior probabilities.

- Most confident selection: select samples that have the highest posterior probability based on one classifier, and at the same time there is certain posterior probability difference between the two classifiers.

- Most different selection: select samples that have the most difference between the two classifiers' posterior probabilities.

The first strategy is appropriate for base classifiers that lack the capability of estimating the posterior probability of their predictions. The second is appropriate for base classifiers that have high classification accuracy and also with high posterior probability. The last one is also appropriate for accurate classifiers and expected to converge faster since big mistakes of one of the two classifiers can be fixed. These sample selection strategies share some similarity with those in previous work (Steedman et al., 2003).

	utter.	**word**	**syll**	**Speaker**
Test Set	102	5,448	8,962	f1a, m1b
Development Set	20	1,356	2,275	f2b, f3b
Labeled set L	5	347	573	m2b, m3b
Unlabeled set U	1,027	77,207	129,305	m4b

Table 1: Training and test sets.

6 Experiments and results

Our goal is to determine whether the co-training algorithm described above could successfully use the unlabeled data for prosodic event detection. In our experiment, 268 ToBI labeled utterances and 886 unlabeled utterances in BU corpus were used. Among labeled data, 102 utterances of all *f1a* and *m1b* speakers are used for testing, 20 utterances randomly chosen from *f2b*, *f3b*, *m2b*, *m3b*, and *m4b* are used as development set to optimize parameters such as λ and confidence level threshold, 5 utterances are used as the initial training set L, and the rest of the data is used as unlabeled set U, which has 1027 unlabeled utterances (we removed the human labels for co-training experiments). The detailed training and test setting is shown in Table 1.

First of all, we compare the learning curves using our proposed confidence-based method to assign possible labels with the simple agreement-based random selection method. We expect that if self-labeling is accurate, adding new samples randomly drawn from these self-labeled data generally should not make performance worse. For this experiment, in every iteration, we randomly select the self-labeled samples that have at least 0.1 difference between two classifiers' posterior probabilities. The number of new samples added to training is 5% of the size of the previous training data. Figure 3 shows the learning curves for accent detection. The number of samples in the x-axis is the number of syllables. The F-measure score using the initial training data is 0.69. The dark solid line in Figure 3 is the learning curve of the supervised method when varying the size of the training data. Compared with supervised method, our proposed relative confidence-based labeling method shows better performance when there is

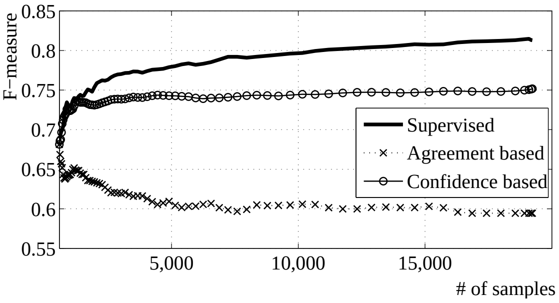

Figure 3: The learning curve of agreement-based and our proposed confidence-based random selection methods for accent detection.

		Confidence	Agreement
Accent detection	% of *P* samples	47%	38%
	P sample error	0.17	0.09
	N sample error	0.12	0.22
IPB detection	% of *P* samples	46%	19%
	P sample error	0.12	0.01
	N sample error	0.18	0.53
Break detection	% of *P* samples	50%	25%
	P sample error	0.15	0.03
	N sample error	0.17	0.42

Table 2: Percentage of positive samples, and averaged error rate for positive (*P*) and negative (*N*) samples for the first 20 iterations using the agreement-based and our confidence labeling methods.

less data, but after some iteration, the performance is saturated earlier. However, the agreement-based method does not yield any performance gain, instead, its performance is much worse after some iteration. The other two prosodic event detection tasks also show similar patterns.

To analyze the reason for this performance degradation using the agreement-based method, we compare the labels of the newly added samples in random selection with the reference annotation. Table 2 shows the percentage of the positive samples added for the first 20 iterations, and the average labeling error rate of those samples for the self-labeled positive and negative classes for two methods. The agreement-based random selection added more negative samples that also have higher error rate than the positive samples. Adding these samples has a negative impact on the classifier's performance. In contrast, our confidence-based approach balances the number of positive and negative samples and significantly reduces the error rates for the negative samples as well, thus leading to performance improvement.

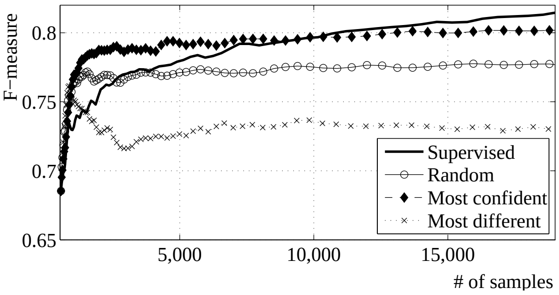

Figure 4: The learning curve of 3 sample selection methods for accent detection.

Next we evaluate the efficacy of the three sample selection methods described in Section 5, namely, random, most confident, and most different selections. Figure 4 shows the learning curves for the three selection methods for accent detection. The same configuration is used as in the previous experiment, i.e., at least 0.1 posterior probability difference between the two classifiers, and adding 5% of new samples in each iteration. All of these sample selection approaches use the confidence-based labeling. For comparison, Figure 4 also shows the learning curve for supervised learning when varying the training size. We can see from the figure that compared to random selection, the most confident selection method shows similar performance in the first few iterations, but its performance continues to increase and the saturation point is much later than random selection. Unlike the other two sample selection methods, most different selection results in noticeable performance degradation after some iteration. This difference is caused by the high self-labeling error rate of selected samples. Both random and most confident selections perform better than supervised learning at the first few iterations. This is because the new samples added have different posterior probabilities by the two classifiers, and thus one of the classifiers benefits from these samples.

Learning curves for the other two tasks (break index and IPB detection) show similar pattern for the random and most different selection methods, but some differences in the most confident selection results. For the IPB task, the learning curve of the most confident selection fluctuates somewhat in the middle of the iterations with similar performance to random selection, however, afterward the performance is better than random selection.

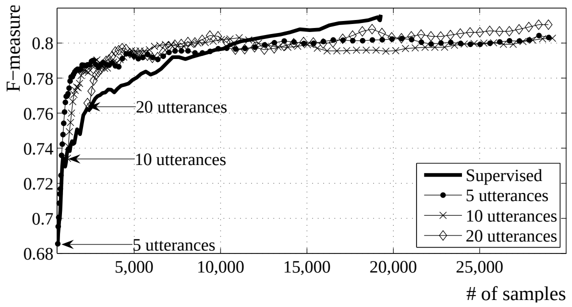

Figure 5: The learning curves for accent detection using different amounts of initial labeled training data.

For the break index detection, the learning curve of most different selection increases more slowly than random selection at the beginning, but the saturation point is much later and therefore outperforms the random selection at the later iterations.

We also evaluated the effect of the amount of initial labeled training data. In this experiment, most confident selection is used, and the other configurations are the same as the previous experiment. The learning curve for accent detection is shown in Figure 5 using different numbers of utterances in the initial training data. The arrow marks indicate the start position of each learning curve. As we can see, the learning curve when using 20 utterances is slightly better than the others, but there is no significant performance gain according to the size of initial labeled training data.

Finally we compared our co-training performance with supervised learning. For supervised learning, all labeled utterances except for the test set are used for training. We used most confident selection with proposed self-labeling method. The initial training data in co-training is 3% of that used for supervised learning. After 74 iterations, the size of samples of co-training is similar to that in the supervised method. Table 3 presents the results of three prosodic event detection tasks. We can see that the performance of co-training for these three tasks is slightly worse than supervised learning using all the labeled data, but is significantly better than the original performance using 3% of hand labeled data.

Most of the previous work for prosodic event detection reported their results using classification accuracy instead of F-measure. Therefore to better compare with previous work, we present below the accuracy results in our approach. The co-training algorithm achieves the accuracy of 85.3%, 90.1%, and 86.7% respectively for accent, intonational phrase boundary, and break index detection, compared with 87.6%, 92.3%, and 88.9% in supervised learning. Although the test condition is different, our result is significantly better than that of other semi-supervised approaches of previous work and comparable with supervised approaches.

		Accent	IPB	Break
Supervised		0.82	0.74	0.77
Co-training	Initial training (3%)	0.69	0.59	0.62
	After 74 iterations	0.80	0.71	0.75

Table 3: The results (F-measure) of prosodic event detection for supervised and co-training approaches.

7 Conclusions

In this paper, we exploit the co-training method for automatic prosodic event detection. We introduced a confidence-based method to assign possible labels to unlabeled data and evaluated the performance combined with informative sample selection methods. Our experimental results using co-training are significantly better than the original supervised results using the small amount of training data, and closer to that using supervised learning with a large amount of data. This suggests that the use of unlabeled data can lead to significant improvement for prosodic event detection.

In our experiment, we used some labeled data as development set to estimate some parameters. For the future work, we will perform analysis of loss function of each classifier in order to estimate parameters without labeled development data. In addition, we plan to compare this to other semi-supervised learning techniques such as active learning. We also plan to use this algorithm to annotate different types of data, such as spontaneous speech, and incorporate prosodic events in spoken language applications.

Acknowledgments

This work is supported by DARPA under Contract No. HR0011-06-C-0023. Distribution is unlimited.

References

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. *Proceedings of* *the Workshop on Computational Learning Theory*, pp. 92-100.

C. W. Wightman and M. Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, Vol. 2(4), pp. 69-481.

G. Levow. 2006. Unsupervised and semi-supervised learning of tone and pitch accent. *Proceedings of HLT-NAACL*, pp. 224-231.

I. Muslea, S. Minton and C. Knoblock. 2000. Selective sampling with redundant views. *Proceedings of the 7th International Conference on Artificial Intelligence*, pp. 621-626.

J. Jeon and Y. Liu. 2009. Automatic prosodic event detection using syllable-base acoustic and syntactic features. *Proceeding of ICASSP*, pp. 4565-4568.

K. Chen, M. Hasegawa-Johnson, and A. Cohen. 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic prosodic model. *Proceedings of ICASSP*, pp. 509-512.

K. Nigam and R. Ghani. 2000 Analyzing the effectiveness and applicability of Co-training *Proceedings 9th International Conference on Information and Knowledge Management*, pp. 86-93.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling English prosody. *Proceedings of ICSLP*, pp. 867-870.

M. Steedman, S. Baker, S. Clark, J. Crim, J. Hockenmaier, R. Hwa, M. Osborne, P. Ruhlen, A. Sarkar 2003. *CLSP WS-02 Final Report: Semi-Supervised Training for Statistical Parsing*.

M. Ostendorf, P. J. Price and S. Shattuck-Hunfnagel. 1995. The Boston University Radio News Corpus. *Linguistic Data Consortium*.

S. Ananthakrishnan and S. Narayanan. 2006. Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling. *Proceedings of ICSLP*, pp. 297-300.

S. Ananthakrishnan and S. Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical and syntactic evidence. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16(1), pp. 216-228.

S. Clark, J. Currant, and M. Osborne. 2003. Bootstrapping POS taggers using unlabeled data. *Proceedings of CoNLL*, pp. 49-55.

S. Dasupta, M. L. Littman, and D. McAllester. 2001. PAC generalization bounds for co-training. *Advances in Neural Information Processing Systems*, Vol. 14, pp. 375-382.

S. Goldman and Y. Zhou. 2000. Enhancing supervised learning with unlabeled data. *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 327-334.

V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan. 2008. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language processing*, pp. 797-811.

W. Wang, Z. Huang, and M. Harper. 2007. Semi-supervised learning for part-of-speech tagging of Mandarin transcribed speech. *Proceeding of ICASSP*, pp. 137-140.