AGREEMENT LEARNING FOR AUTOMATIC ACCENT ANNOTATION

Xinqiang Ni[*1] *Yining Chen*[2] *Min Chu*[2] *Frank K. Soong*[2] *Yong Zhao*[2] *Ping Zhang*[1]

[1]Institute of Electronics, Chinese Academy of Sciences, Beijing, China
[2]Microsoft Research Asia, Beijing, China
[1]xqni@mails.gucas.ac.cn, [2]{ynchen, minchu, frankkps, yzhao}@microsoft.com,
[1]pzhang@mail.ie.ac.cn

ABSTRACT

Automatic accent annotation is important in both speech synthesis and speech recognition. Existing statistical learning algorithms rely heavily on a sufficiently large set of labeled training samples that are expensive and time consuming to collect. For unlabeled data, unsupervised learning can be initiated with a small set of manually labeled data. This paper shows that the accuracy of automatic accent annotation can be improved by augmenting a small amount of manually labeled data with a large pool of unlabeled data. We introduce an agreement-learning algorithm for this propose. Experimental results show that it is possible to reduce human-labeling effort significantly while reducing up to 50% errors.

Index Terms— Agreement learning, Semi-supervised learning, Accent detection

1. INTRODUCTION

Prosody labeling is important for both speech synthesis and automatic speech understanding. Among all prosody events, accent is probably the most prominent. Manually labeling accent is quite time consuming. This paper will focus on accent auto-labeling.

There are some existing statistical learning algorithms for annotating accent [1-9]. One key difficulty with these current algorithms, and the principle issue addressed by this paper, is that a large number of labeled training samples are required for accurate machine learning. However, labeled instances are difficult and time consuming to obtain, since they require experienced human annotators. Meanwhile, in text-to-speech synthesis systems, there is a far greater amount of unlabeled data available than the labeled data. Recently, some work relating to how to use unlabeled data has been done [1-3]. Unsupervised learning algorithms like k-means are used in [3]. In [1], a semi-supervised learning algorithm is adopted.

Semi-supervised learning is a commonly employed way to exploit unlabeled data. It uses unlabeled data to improve models. There are many kinds of semi-supervised learning algorithms, They include self-training [10], co-training [11], and graph-based methods [12]. "Self-training" is a traditional method and is used in some speech applications [1, 2, 4]. It automatically labels unlabeled samples by using a small number of human labeled samples as "seed" samples, and re-trains the model with the auto-labeled data.

In this paper, we propose a new algorithm we call agreement learning. In this algorithm, ensembles of different classifiers are combined to make a decision on the unlabeled data. In each iteration, only unlabeled data got same labels in each classifier are used as training data in the new model. However, this kind of combination often acquires classifiers with different output but similar performance. In the accent annotation task, acoustic classifier and linguistic classifiers are these kinds of classifiers.

Acoustic cues, such as intensity, duration, and fundamental frequency are used to develop predictions of accent for a given utterance [4, 5]. Spectral parameters such as Mel-scale Frequency Cepstral Coefficients (MFCC) are used in some accent detection studies [6]. Linguistic cues derived from texts, such as part of speech (POS), N-Grams of POS, and the positions within the phrase are used in accent detection as well [7-9]. The classifiers for these two cues are independent of each other and experiments show that the performance is similar [2]. We find the ensemble from these two classifiers can help a lot.

In Section 2, semi-supervised learning is introduced. In Section 3, agreement learning is described. Evaluations and results are presented in Section 4 and conclusions are outlined in Section 5.

2. SEMI-SUPERVISED LEARNING

Unlabeled data are generally not sufficient to train a classifier for better-than-random classification performance [13]. However, even without class labels, such data still provide some information on the joint distribution of

[*] This work is done when the first author visits Microsoft Research Asia as an intern

features.

Supervised learning requires labeled training data to train reasonable classifiers while unsupervised learning is employed to discover hidden structure in unlabeled data. Semi-supervised learning is between supervised and unsupervised learning and requires only a small amount of labeled training data. It improves performance using additional unlabeled data.

The basic idea behind semi-supervised learning is to automatically label unlabeled samples by using a small number of human labeled data so-called seeds. By doing this, semi-supervised learning yields a larger labeled dataset that can be used as training data for supervised learning. The aim of semi-supervised training is to exploit unlabeled data to improve the performance of a classifier. In doing so, semi-supervised training uses unlabeled data to modify hypotheses obtained from stand-alone labeled data.

Self-learning has been used in speech recognition [14] and prosody auto-labeling [2]. It is summarized in Figure 1. First, we create an initial acoustic classifier based solely on labeled data. Then, a two-step procedure is performed: first, an acoustic classifier is used to label all unlabeled data; then, a new acoustic classifier based on all the data is learned. Intuitively, self-learning tries to find the most likely hypothesis that could generate the unlabeled data distribution. Self-learning can be seen as clustering of unlabeled data "around" the samples in the original training set.

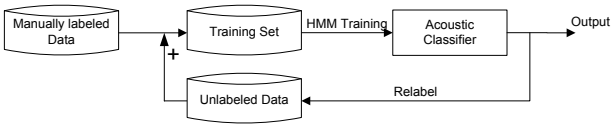

Figure 1: Semi-supervised learning.

Since only one classifier is used, the classifier will cause errors to unlabeled data during each iteration. Then, the classifier is re-trained with those data that contains errors. These errors will come back to the models. Some errors will always be there, And to avoid this behavior, we introduce agreement learning.

3. AGREEMENT LEARNING

We present an algorithm named agreement learning: the samples which are in agreement among different classifiers can be exploited using semi-supervised learning methods.

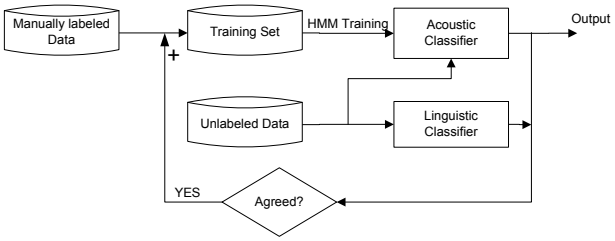

Figure 2: Agreement learning.

A flowchart is depicted in Figure 2. First, we train an initial acoustic model using the labeled data, and then classify the unlabeled ones. Next, we compare labels assigned by the acoustic classifier and linguistic classifier. If they agree, we add data to the training set. Then, a new acoustic classifier is trained with manually labeled data and selected machine-labeled data. The last two steps are iterated, and in each pass, as the more valuable data are added to the training set, the acoustic model is more accurate. Now we have two classifiers: a Hidden Markov Model (HMM) based acoustic classifier and a linguistic classifier. The HMM-based acoustic classifier aims at exploiting the segmental information of accent vowels. The linguistic classifier captures the text level information. The two classifiers are introduced below.

Some other algorithms employ similar ideas. In speech recognition, people combine different decoder by cross adaptation.

3.1. Linguistic classifier

According to Pike [15], usually content words, which carry more semantic weight in a sentence, are accented while function words are unaccented. Following this rule, a simple linguistic classifier is designed in this study: according to their POS tags, content words are deemed as accented while non-content or function words are deemed as unaccented. In our study, nouns, verbs, adjectives, and adverbs are content words and others words are function words. The accent detection accuracy of content words is as high as 96%. And for function words, it is about 85%. Since the self-consistency of labelers for content words is about 97%, we can assume the accuracy in content word is high enough. Function words are the only part we need to label accent.

3.2. HMM-based acoustic classifier

The HMM-based acoustic classifier uses segmental information that can distinguish accented vowels from unaccented ones.

First, the pronunciation lexicon is adjusted in terms of the accent- and position-dependent phone set. Each word pronunciation is encoded into both accented and unaccented versions. In the accented one, the vowel in the primary stress

syllable is accented and all the other vowels are unaccented. In the unaccented word, all vowels are unaccented.

In the training process, the phonetic transcription of the accented version of a word is used if it is accented. Otherwise, the unaccented version is used. In addition to the above adjustment, the whole training process is the same as conventional speech recognition training. Accent- and position-dependent HMM are trained with the standard Baum-Welch algorithm in the HTK software package [16].

In the decoding part, the trained acoustic model is used to label accent. Given an unknown utterance the most likely path is found.

Linguistic classifier is based on syntactic cues and acoustic classifier is based on acoustic cues. Since the two classifiers are independent and generate accent labels from different information sources, they do not always agree with each other. We can consider that the agreed labels are more accurate and more suitable for self-learning. This is the main topic of our agreement method.

4. EXPERIMENTS AND RESULTS

4.1. Experiment settings

The speech corpus evaluated consists of 6,412 utterances by a professional female broadcaster. We used 500 utterances as the labeled set for initial acoustic training and 5,412 utterances as the unlabeled set for semi-supervised learning, including 500 utterances as the test set. In these experiments, only the error rates among function words are studied.

4.2. Accuracy of four different methods

First we compare the accuracy of four different methods: the acoustic classifier, the linguistic classifier, self-learning, and agreement learning. The results are shown in Figure 3. It is evident that agreement learning is the best. When using it, the accuracy increases to 91.28%. Compared with the baseline of 84.30% provided by the linguistic classifier, a 45% relative reduction in the error rate is achieved.

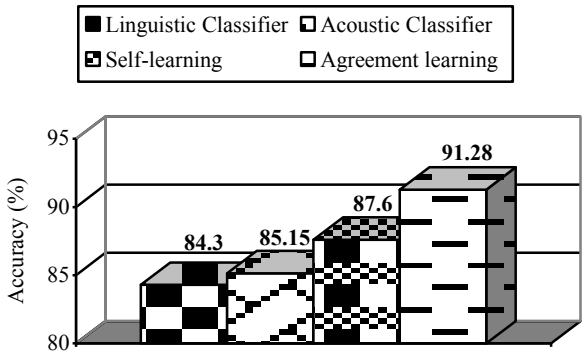

Figure 3: Accuracy of four different methods.

4.3. Comparing self-learning with agreement learning

In this experiment, we hold the number of unlabeled data constant and vary the number of labeled data, and then compare the accuracy of self-learning with agreement learning. Figure 4 shows the performance of the two different methods. The vertical axis indicates accuracy on test sets, and the horizontal axis indicates the amount of labeled training data.

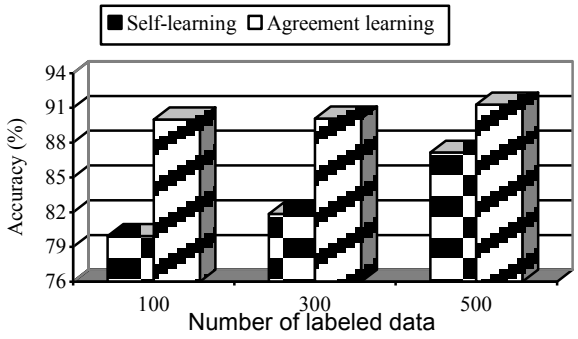

Figure 4: Performance of semi-supervised learning and agreement learning.

Figure 4 shows that agreement learning achieves encouraging improvement over self-learning in all training sets. For example, when using 100 labeled data, accuracy increases from 79.84% to 90.01%. This represent a 50% reduction in errors. Also, we can find agreement learning did less well when the size of manual labeled data increased. When the size of manual labeled data is small, the performance of the acoustic classifier is not optimum. In this condition, there will be noise in the self-learning training data. Then it does not improve the performance by much. On the other hand, agreement learning, which includes only agreement data, is less affected.

4.4. Performance of varying the number of unlabeled data

In Figure 5, we consider the effects of varying the amount of unlabeled data using agreement learning. The vertical axis indicates accuracy on test sets, and the horizontal axis indicates the amount of unlabeled data. For three different quantities of labeled data, we hold the number of labeled data constant, and vary the number of unlabeled data. Notice that there will be an obvious improvement in the accuracy. For example, with 100 labeled data, when we increase the number of unlabeled data from 1,000 to 5,000, classification accuracy increases from 89.04% to 90.01%.

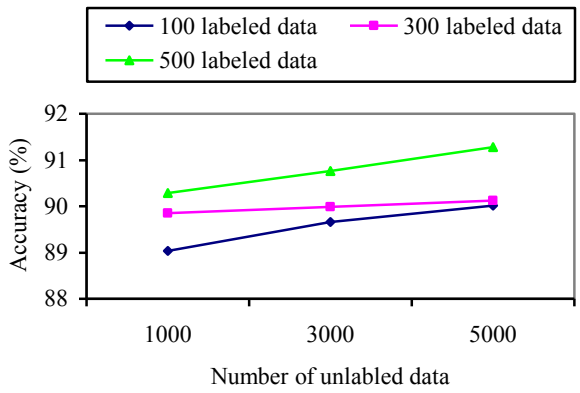

Figure 5: Accuracy while varying the number of unlabeled data.

These experimental results demonstrate that agreement learning can improve classification and reduce the need for manual labels.

5. CONCLUSIONS AND DISCUSSION

This paper describes an agreement learning algorithm for automatic accent annotation. This method combines two complementary classifiers in semi-supervised learning. In each iteration of semi-supervised learning, only the samples which get the same result between two classifiers are used to train the new classifier.

Our experiments on a large speech corpus demonstrate that accuracy can be significantly improved by using unlabeled data. With 500 labeled data, a traditional acoustic classifier reaches 85.15% accuracy while self-learning achieves 87.15%. While using agreement learning, the accuracy increases to 91.28%, which represents a 41.28% reduction in errors. Agreement learning can cut the manual labeling effort dramatically.

In future work, other prosody events such as break and boundary tone will be labeled together with accent.

6. ACKNOWLEDGMENTS

The authors would like to thank Scott Meredith for his great help on creating the specification for prosody annotation. We would also like to offer special thanks to Yaya Peng for creating these accent labels.

7. REFERENCES

[1] Y.N. Chen, M. Lai, M. Chu, etc, "Automatic Accent Annotation with Limited Manually Labeled Data", in Proc. of Speech Prosody, 2006.

[2] M. Lai, Y.N. Chen, M. Chu, etc, "A Hierarchical Approach to Detect Stress in English Sentences", in Proc. of ICASSP, pp 753-756, 2006.

[3] S. Ananthakrishnan, S. Narayanan, "Combining Acoustic, Lexical, and Syntactic Evidence for Automatic Unsupervised Prosody Labeling", in Proc. of InterSpeech 2006, pp 297-300, 2006.

[4] C.W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns", IEEE Trans. on Speech and Audio Processing, 2(4), pp 469-481, 1994.

[5] I. Bulyko and M. Ostendorf. "A Bootstrapping Approach to Automating Prosodic Annotation for Constrained Domain Synthesis", in Proc. of the IEEE Workshop on Speech Synthesis, pp 115-118, 2002.

[6] A. Conkie, G. Riccardi, and R.C. Rose "Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events", in Proc. of Eurospeech, pp 523-526, 1999.

[7] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara, and M. Dantsuji, "Modeling and automatic detection of English sentences accent for computer-assisted English prosody learning system", in Proc. of ICSLP, pp 749-752, 2002.

[8] K. Chen, and M. Hasegawa-Johnson, "An automatic prosody labeling system using ANN-based syntacticprosodic model and GMM-based acoustic-prosodic model", in Proc. of ICASSP, pp 509-512, 2004.

[9] S. Arnfield, "Prosody and syntax in corpus based analysis of spoken English," Ph.D. dissertation, University of Leeds, Dec. 1994.

[10] S. Clark, J.R. Curran, and M. Osborne, "Bootstrapping POS taggers using unlabelled data". In Walter Daelemans and Miles Osborne, editors, Proceedings of CoNLL, pp 49-55, 2003.

[11] A, Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training", In Proc. of 11th Annual Conf. on Comp. Learning Theory (COLT), pp 92–100, 1998.

[12] D.A. Zighed, S. Lallich, and F. Muhlenbach, "Separability Index in Supervised Learning", In Principles of Data Mining and Knowledge Discovery, Proc. Of the 6th European Conference PKDD, pp475-487, 2002.

[13] V. Castelli, T.M. Cover, "On the exponential value of labeled samples", Pattern Recognition Letters, 16(1), pp105-111, 1995.

[14] G. Tur, D. Hakkani-Tur, A. Chotimongkol, "Semi-Supervised Learning for Spoken Language Understanding Using Semantic Role Labeling", in Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp 319-324, 2005.

[15] E.C. Kuhlen, "An Introduction to English Prosody", Edward Arnold, 1986.

[16] S. Young, G. Evermann, D. Kershaw, etc, "HTK Book, version 3.1", http://htk.eng.cam.ac.uk/protdocs/htk_book.shtml