Active Learning for Classifying Phone Sequences from Unsupervised Phonotactic Models

Shona Douglas
AT&T Labs - Research
Florham Park, NJ 07932, USA
shona@research.att.com

Abstract

This paper describes an application of active learning methods to the classification of phone strings recognized using unsupervised phonotactic models. The only training data required for classification using these recognition methods is assigning class labels to the audio files. The work described here demonstrates that substantial savings in this effort can be obtained by actively selecting examples to be labeled using confidence scores from the BoosTexter classifier. The saving in class labeling effort is evaluated on two different spoken language system domains in terms both of the number of utterances to be labeled and the length of the labeled utterances in phones. We show that savings in labeling effort of around 30% can be obtained using active selection of examples.

1 Introduction

A major barrier to the rapid and cost-effective development of spoken language processing applications is the need for time-consuming and expensive human transcription and annotation of collected data. Extensive transcription of audio is generally undertaken to provide word-level labeling to train recognition models. Applications that use statistically trained classification as a component of an understanding system also require this transcribed text to train on, plus an assignment of class labels to each utterance.

In recent work by Alshawi (2003) reported in this conference, new methods for unsupervised training of phone string recognizers have been developed, removing the need for word-level transcription. The phone-string output of such recognizers has been used in classification tasks using the BoosTexter text classification algorithm, giving utterance classfication accuracy that is surprisingly close to that obtained using conventionally trained word trigram models requiring transcription. The only training data required for classification using these recognition methods is assigning class labels to the audio files. The aim of the work described in this paper is to amplify this advantage by reducing the amount of effort required to train classifiers for phone-based systems by actively selecting which utterances to assign class labels. Active learning has been applied to classification problems before (McCallum and Nigam, 1998; Tur et al., 2003), but not to classifiying phone strings.

2 Unsupervised Phone Recognition

Unsupervised recognition of phone sequences is carried out according to the method described by Alshawi (2003). In this method, the training inputs to recognition model training are simply the set of audio files that have been recorded from the application.

The recognition training phase is an iterative procedure in which a phone n-gram model is refined successively: The phone strings resulting from the current pass over the speech files are used to construct the phone n-gram model for the next iteration. We currently only re-estimate the n-gram model, so the same general-purpose HMM acoustic model is used for ASR decoding in all iterations.

Recognition training can be briefly described as follows. First, set the phone sequence model to an initial phone string model. This initial model used can be an unweighted phone loop or a general purpose phonotactic model for the language being recognized. Then, for successively larger n-grams, produce the output set of phone sequences from recognizing the training speech files with the current phone sequence model, and train the next larger n-gram phone sequence model on this output corpus.

3 Training phone sequence classifiers with active selection of examples

The method we use for training the phone sequence classifier is as follows.

1. Choose an initial subset S of training recordings at random; assign class label(s) to each example.

2. Recognize these recordings using the phone recognizer described in section 2.

3. Train an initial classifier C on the pairs (*phone string*, *class label*) of S.

4. Run the classifier on the recognized phone strings of the training corpus, obtaining confidence scores for each classification.

5. While labeling effort is available, or until performance on a development corpus reaches some threshold,

   (a) Choose the next subset S' of examples from of the training corpus, on the basis of the confidence scores or other indicators. (Selection criteria are discussed later.)

   (b) Assign class label(s) to each selected example.

   (c) Train classifier C' on all the data labeled so far.

   (d) Run C' on the whole training corpus, obtaining confidence scores for each classification.

   (e) Optionally test C' on a separate test corpus.

4 Experimental Setup

The datasets tested on and the classifier used are the same as those in the experiments on phone sequence classification reported by Alshawi (2003). The details are briefly restated here.

4.1 Data

Two collections of utterances from two domains were used in the experiments:

1. *Customer care* utterances (HMIHY). These utterances are the customer side of live English conversations between AT&T residential customers and an automated customer care system. This system is open to the public so the number of speakers is large (several thousand).

The total number of training utterances was 40,106. All tests use 9724 test utterances. Average utterance length was 11.19 words; there were 56 classes, with an average of 1.09 classes per utterance.

2. *Text-to-Speech Help Desk* utterances (TTSHD). This is a smaller database of utterances in which customers called an automated information system primarily to find out about AT&T Natural Voices text-to-speech synthesis products.

The total number of possible training utterances was 10,470. All tests use 5005 test utterances. Average utterance length was 3.95 words; there were 54 classes, with an average of 1.23 classes per utterance.

4.2 Phone sequences

The phone sequences used for testing and training are those obtained using the phone recognizer described in section 2. Since the phone recognizer is trained without labeling of any sort, we can use all available training utterances to train it, that is, 40,106 in the HMIHY domain and 10,470 in the TTSHD domain. The initial model used to start the iteration is, as in (Alshawi, 2003), an unweighted phone loop.

4.3 Classifier

For the experiments reported here we use the BoosTexter classifier (Schapire and Singer, 2000). The features used were identifiers corresponding to prompts, and phone n-grams up to length 4. Following Schapire and Singer (2000), the confidence level for a given prediction is taken to be the difference between the scores assigned by BoosTexter to the highest ranked action (the predicted action) and the next highest ranked action.

4.4 Selection criteria

Subsets of the recognized phone sequences were selected to be assigned class labels and used in training the classifiers. Examples were selected in order of BoosTexter confidence score, least confident first. Further selection by utterance length was also used in some experiments such that only recognized utterances with less than a given number of phones were selected.

5 Experiments

5.1 Evaluation metrics

We are interested in comparing the performance for a given amount of labeling effort of classifiers trained on random selection of examples with that of classifiers trained on examples chosen according to the confidence-based method described in section 3.

The basic measurements are:

$A(e)$: the classification accuracy at a given labeling effort level e of the classifier trained on actively selected labeling examples.

$R(e)$: the classification accuracy at a given labeling effort level e of the classifier trained on randomly selected labeling examples.

$A^{-1}(R(e))$: the effort required to achieve the performance of random selection at effort e, using active learning.

Derived from these is the main comparison we are interested in:

Effort	A	R	$A^{-1}(R)$	*Effort*
(utt)	*(%)*	*(%)*	*(utt)*	*Ratio*
2000	67.4	66.0	1128	0.56
4000	69.6	68.0	2678	0.67

Table 1: HMIHY, no length limit, effort is number of utterances

Effort	A	R	$A^{-1}(R)$	*Effort*
(phn)	*(%)*	*(%)*	*(phn)*	*Ratio*
68032	67.0	66.1	52940	0.78
128636	69.3	67.9	91057	0.71

Table 2: HMIHY, length limited, effort is number of phones

$\textit{EffortRatio}(e) = A^{-1}(R(e))/e$: the proportion of the effort that would be required to achieve the performance of random selection at effort e, actually required using active learning: that is, low is good.

We use two metrics for labeling effort: the number of utterances to be labeled and the number of phones in those utterances. The number of phones is indicative of the length of the audio file that must be listened to in order to make the class label assignment, so this is relevant to assessing just how much real effort is saved by any active learning technique.

5.2 Results

Table 1 gives the results for selected levels of labeling effort in the HMIHY domain, calculated in terms of number of utterances labeled.

These results suggest that we can achieve the same accuracy as random labeling with around 60% of the effort by active selection of examples according to the confidence-based method described in section 3.

However, a closer inspection of the chosen examples reveals that, on average, the actively selected utterances are nearly 1.5 times longer than the random selection in terms of number of phones. (This is not suprising given that the classification method performs much worse on longer utterances, and the confidence levels reflect this.) In order to overcome this we introduce as part of the selection criteria a length limit of 50 phones. This allows us to retain appreciable effort savings as shown in table 2.

The TTSHD application is considerably less complex than HMIHY, and this may be reflected in the greater savings obtained using active learning. Tables 3 and 4 show the corresponding results for this domain.

There is also a smaller variation in utterance length between actively and randomly selected training examples (more like 110% than the 150% for HMIHY); table 4 shows that defining effort in terms of number of phones still results in appreciable savings for active learning. (Incorporating a length limit gave little additional benefit here.)

Effort	A	R	$A^{-1}(R)$	*Effort*
(utt)	*(%)*	*(%)*	*(utt)*	*Ratio*
2000	78.9	77.5	1327	0.66
4000	80.3	78.8	1971	0.49

Table 3: TTSHD, effort is number of utterances

Effort	A	R	$A^{-1}(R)$	*Effort*
(phn)	*(%)*	*(%)*	*(phn)*	*Ratio*
35877	78.9	77.9	27019	0.75
71338	80.3	79.1	48267	0.68

Table 4: TTSHD, effort is number of phones

6 Discussion

By actively choosing the examples with the lowest confidence scores first, we can get the same classification results with around 60-70% of the utterances labeled in HMIHY and TTSHD. But we want to optimize labeling effort, which is presumably some combination of a fixed amount of effort per utterance plus a “listening effort” proportional to utterance length. We therefore augmented our active learning selection to include a constraint on the length of the utterances, measured in recognized phones. If we simply take effort to be proportional to the number of phones in the utterances selected (likely to result in a conservative estimate of savings), the effort reduction at 4,000 utterances is around 30% even for the more complex HMIHY domain. Further investigation is needed into the best way to measure overall labeling effort, and into refinements of the active learning process to optimize that labeling effort.

References

H. Alshawi. 2003. Effective utterance classification with unsupervised phonotactic models. In *HLT-NAACL 2003*, Edmonton, Canada.

A. K. McCallum and K. Nigam. 1998. Employing EM in pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning*, pages 350–358.

R. E. Schapire and Y. Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

Gokhan Tur, Robert E. Schapire, , and Dilek Hakkani-Tur. 2003. Active learning for spoken language understanding. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP’03)*, Hong Kong, April. (to appear).