Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2009) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion

Dong Yu [a,*,1], Balakrishnan Varadarajan [b,2], Li Deng [a], Alex Acero [a]

[a] *Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA*
[b] *Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA*

Received 24 September 2008; received in revised form 15 December 2008; accepted 11 March 2009

Abstract

We propose a unified global entropy reduction maximization (GERM) framework for active learning and semi-supervised learning for speech recognition. Active learning aims to select a limited subset of utterances for transcribing from a large amount of un-transcribed utterances, while semi-supervised learning addresses the problem of selecting right transcriptions for un-transcribed utterances, so that the accuracy of the automatic speech recognition system can be maximized. We show that both the traditional confidence-based active learning and semi-supervised learning approaches can be improved by maximizing the lattice entropy reduction over the whole dataset. We introduce our criterion and framework, show how the criterion can be simplified and approximated, and describe how these approaches can be combined. We demonstrate the effectiveness of our new framework and algorithm with directory assistance data collected under the real usage scenarios and show that our GERM based active learning and semi-supervised learning algorithms consistently outperform the confidence-based counterparts by a significant margin. Using our new active learning algorithm cuts the number of utterances needed for transcribing by 50% to achieve the same recognition accuracy obtained using the confidence-based active learning approach, and by 60% compared to the random sampling approach. Using our new semi-supervised algorithm we can determine the cutoff point in determining which utterance-transcription pair to use in a principled way by demonstrating that the point it finds is very close to the achievable peak point.

Keywords: Active learning; Semi-supervised learning; Acoustic model; Entropy reduction; Confidence; Lattice; Collective information

* Corresponding author. Tel.: +1 425 707 9282; fax: +1 425 823 8659.
E-mail addresses: dongyu@microsoft.com (D. Yu), bvarada2@jhu.edu (B. Varadarajan), deng@microsoft.com (L. Deng), alexac@microsoft.com (A. Acero).
[1] Some preliminary results will be presented at ICASSP conference to be held in Taipei, Taiwan, 2009.
[2] This work was carried out during the internship program at Microsoft research.

0885-2308/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2009.03.004

1. Introduction

In the recent years, we have witnessed great progress in deploying real world interactive voice response (IVR) systems. A typical example of these real world systems is the voice search application (Yu et al., 2007), with which users may search for information such as phone number of a business with voice. There are two key differentiators of these systems to the earlier IVR systems. First, the vocabulary size of these systems is usually large, typically over 10 K. Second, users often interact with the system using free-style instantaneous speech under real noisy environments. These differences post great challenges in promoting the system's performance to a high level. In these systems, getting un-transcribed data is usually as cheap as logging the users' interactions with the system, while getting transcribed data can be very costly.

In this paper we investigate approaches to improving the automatic speech recognition (ASR) systems' performance in the applications where the initial accuracy is very low and only small amount of data can be transcribed. We tackle the problem with active learning and semi-supervised learning approaches and propose to unify these two approaches under the global entropy reduction maximization (GERM) framework.

The concept of active learning has been proposed and studied in the machine learning community for many years (Cohn et al., 1994; Anderson and Moore, 2005; Anderson et al., 2006; Ji et al., 2006) and has been applied to the development of spoken dialog systems (Hakkani-Tr et al., 2004; Riccardi and Hakkani-Tur, 2005; Kuo and Goel, 2005; Tur et al., 2005) and acoustic models (Kamm and Meyer, 2003; Kamm and Meyer, 2004; Kamm, 2004; Hakkani-Tur and Gorin, 2002) in the recent several years. The basic idea of active learning is to actively ask a question based on all the information available so far, so that some objective function can be optimized when the answer becomes known. In many tasks (e.g., improving dialog systems and acoustic models) the question to be asked is limited to selecting an utterance for transcribing from a set of un-transcribed utterances.

Four criteria have been proposed in the active learning literature for selecting samples: In the confidence-based approach (Hakkani-Tur and Gorin, 2002; Riccardi and Hakkani-Tur, 2005), samples with the lowest confidence are selected for transcribing. In the query-by-committee based approach (Dagan and Engelson, 1995), samples that cause biggest different opinions from a set of recognizers (committee) are selected. In the confusion (entropy) reduction based approach, samples that reduce the entropy about the true model parameters are selected for transcribing, and in the error rate-based approach (Kuo and Goel, 2005), the samples that can minimize the expected error rate most is selected. The confidence-based approach is the approach used most in the spoken dialog systems (Hakkani-Tr et al., 2004; Riccardi and Hakkani-Tur, 2005; Tur et al., 2005) and acoustic models (Kamm and Meyer, 2003; Kamm and Meyer, 2004; Kamm, 2004; Hakkani-Tur and Gorin, 2002) due to its simplicity and proven effectiveness.

The semi-supervised learning of acoustic models (AMs) has also been studied for many years (Wessel et al., 1998; Kemp and Waibel, 1999; Charlet, 2001; Moreno and Agarwal, 2003; Zhang and Rudnicky, 2006). In the semi-supervised learning, there is a transcribed set and an un-transcribed set. The task is to select the transcriptions automatically for those un-transcribed utterances so that the system trained using the combined data set performs best according to some criterion. Typical approaches used in speech recognition include incremental training where the high-confidence (determined with a threshold) utterances are combined with transcribed utterances (if available) to adapt or retrain the recognizer and then use the adapted recognizer to select the next batch of utterances, and the generalized expectation maximization (GEM) where all utterances are used but with different weights determined by the confidence. Note that both these methods are confidence-based. It has been shown that these approaches have the drawback of reinforcing what the current model already knows and even reinforcing the errors and cause divergence if the performance of the current model is very poor (which is the case in voice search applications).

Note that the confidence-based active learning and semi-supervised learning approaches select the utterances solely based on the confidence of individual utterances. Our framework proposed in this paper differs from these existing approaches in that we make the decision on its effect to the whole dataset. More specifically, our active learning and semi-supervised learning algorithms focus on the improvement to the overall system by taking into consideration the confidence of each utterance, the frequency of the similar and contradictory patterns in the un-transcribed set when selecting the utterances for transcribing or determining the right utterance-transcription pair to be included in the semi-supervised training set. Both these algorithms

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

estimate the expected entropy reduction each utterance or the utterance-transcription pair may cause to the whole un-transcribed dataset and can be unified under the GERM framework. We also show that the active learning and semi-supervised learning approaches can be combined to achieve even better results with the available un-transcribed data set and the amount of data allowed to be transcribed.

We demonstrate the effectiveness of our new framework and algorithm with directory assistance (Yu et al., 2007) data collected under real usage scenarios and show that the GERM based active learning and semi-supervised learning algorithms consistently outperform the confidence-based counterparts by a significant margin. Our new active learning algorithm cuts the number of utterances needed for transcribing by 50% to achieve the same recognition accuracy obtained using the confidence-based approach, and by 60% compared to the random sampling approach. Using our new semi-supervised algorithm we can determine the cut-off point in a principled way.

The organization of the paper is as follows. In Section 2, we introduce our novel active learning algorithm that maximizes the global entropy reduction. We describe the intuition behind our criterion and derive the main formulas associated with the criterion. In Section 3, we describe the semi-supervised algorithm that uses the information in the whole dataset. We illustrate the motivation behind using the collective information in determining the utterance-transcription pairs and show how the criterion can be fit into the GERM framework. In Section 4, a unified framework and associated procedure is given. We analyze the word recognition experiments and results on the directory assistance data in Section 5 providing evidence for the effectiveness of our new techniques, and conclude the paper in Section 6.

2. Active learning with global entropy reduction maximization criterion

Heuristically, transcribing the least confident utterances can provide the most information to the system and this is the reason most existing confidence-based active learning approaches select the utterances that are least confident for transcribing. While this strategy seems to be reasonable it has some limitations. For example, we have observed that the conventional confidence-based active learning algorithm tends to select noise and garbage utterances since these utterances typically have low confidence scores. Unfortunately, transcribing these utterances is usually difficult and carries little value in improving the ASR performance.

The above limitation comes from the fact that the existing confidence-based active learning approaches make the decision based on gains on one utterance only. Transcribing the least confident utterance can greatly help recognizing that utterance. However, it may not be helpful in improving the recognition accuracy on other utterances. Consider two speech utterances A and B where A has a slightly lower confidence score than B has. If A is observed only once and B occurs frequently in the dataset, a reasonable choice is to transcribe B instead of A since transcribing B would correct a larger fraction of errors in the test data than transcribing A and thus has better potential to improve the performance of the whole system. This example shows that we should select the utterances that can provide the most benefit to the whole dataset and this is the core idea of our GERM based active learning algorithm.

We would like to point out that using a global criterion for active learning has also been explored by Kuo and Goel (2005) for the dialog system upon the error rate reduction approaches. Different from their approach, our approach maximizes the expected lattice entropy reduction instead of the error rate over all the un-transcribed data from which we wish to select. Optimizing the entropy is more robust than optimizing the top choice since it considers all possible outcomes weighted with probabilities. Furthermore, Kuo and Goel (2005) focused on the static classification problem which is a much easier problem to work with than the ASR problem on which we focus in this paper. ASR is a sequential recognition problem and we need to consider the segments in the lattices or recognition results when estimating the gains.

Put formally, let $X_1, X_2, \ldots, X_n$ be the n candidate speech utterances and $L_1, L_2, \ldots, L_n$ be the lattices generated by the speech recognizers in response to the utterances $X_1, X_2, \ldots, X_n$ respectively. We wish to choose a subset $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ from these n utterances for transcribing such that the expected reduction of entropy in the lattices $L_1, L_2, \ldots, L_n$ between the original AM Θ and the new model Θ^s over the whole dataset

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

$$E[\Delta H(L_1,\ldots,L_n|X_{i_1},\ldots,X_{i_k})] = \quad (1)$$

$$E[H(L_1,\ldots,L_n|\Theta) - H(L_1,\ldots,L_n|\Theta^s)] = \quad (2)$$

$$E[H(L_1,\ldots,L_n|\Theta)] - E[H(L_1,\ldots,L_n|\Theta^s)] = \quad (3)$$

$$H(L_1,\ldots,L_n|\Theta) - E[H(L_1,\ldots,L_n|\Theta^s)] \quad (4)$$

is maximized. Note that the true transcription T_{i_k} of the utterance X_{i_k} is unknown when we select the utterances and that is the reason we optimize the expected (averaged) value of the entropy reduction over all possible transcriptions.

Note that this optimization problem is expensive to solve since the inclusion of one utterance would affect the selection of another. For example, once an utterance is chosen, the need for selecting utterances that are acoustically similar to the chosen one becomes smaller. To make the problem tractable we approximate the solution to this optimization problem with a greedy search algorithm. We select a single utterance that maximizes the expected entropy reduction over the whole dataset, adjust the entropies for all similar utterances, and determine the next utterance that gives us the highest gain. This process continues until we reach the number of utterances allowable for transcribing.

To optimize the GERM criterion we approximated the expected entropy reduction when an utterance X_i is selected for transcribing as

$$E[\Delta H(L_1,\ldots,L_n|X_i)] \cong \sum_{j=1}^{n} E\left[\Delta H\left(L_j|X_i\right)\right] = \sum_{j=1}^{n} E\left[\Delta H^a_{j|i}\right] \quad (5)$$

where we have assumed that the utterances are independently drawn. The expected entropy reduction over L_j with X_i selected for transcribing $E[\Delta H^a_{j|i}]$ can be estimated with a distance-based approach as

$$E\left[\Delta H^a_{j|i}\right] \cong \alpha H\left(L_j|\Theta\right)\mathrm{e}^{-\beta d\left(X_i,X_j\right)} \quad (6)$$

where α and β are parameters related to the training algorithm used and the number of transcribed utterances in the initial training set and may be estimated from the initial transcribed training set, and $d(X_i,X_j)$ is the distance between the utterances X_i and X_j where $d(X_i,X_j) = 0$ if two utterances are the same and $d(X_i,X_j) = \infty$ if two utterances do not have common phones in the lattices.

Let us examine two extreme cases. If $d(X_i,X_j) = 0$ (e.g., $X_i = X_j$) then the expected entropy reduction on L_j is proportional to its original entropy, or

$$E\left[\Delta H^a_{j|i}\right] \cong \alpha H\left(L_j|\Theta\right). \quad (7)$$

On the other hand, if $d(X_i,X_j) = \infty$, i.e., L_i and L_j does not have common phones, the AM of any of the phones in the lattice L_j will not be updated after the retraining when the utterance X_i is selected for transcribing. This implies that the acoustic scores and hence the probabilities of all the paths in the lattice L_j will remain the same, or

$$E\left[\Delta H^a_{j|i}\right] = 0. \quad (8)$$

The distance $d(X_i,X_j)$ can be estimated in many different ways. For example, we may use the dynamic time warping (DTW) distance between the utterances X_i and X_j as the distance $d(X_i,X_j)$. In this paper we have used the Kullback–Leibler divergence (KLD) between two lattices L_i and L_j as the distance. The reason KLD was used in our study is that we believe the effect of X_i to X_j is different from that of X_j to X_i and it has been proven to be effective in our experiments. For example lattices L_i and L_j both confuse between words star, stark and start with probabilities $P_i(star) = 0.3$, $P_i(stark) = 0.3$, $P_i(start) = 0.4$ and $P_j(star) = 0.4$, $P_j(stark) = 0.4$, $P_j(start) = 0.2$. The initial entropy of lattice L_j is 1.522 nats. The distance between two lattices is estimated as $d(X_i,X_j) = KLD(0.3,0.3,0.4;0.4,0.4,0.2) = 0.3log_2(0.3/0.4) + 0.4log_2(0.3/0.4) + 0.4log_2(0.4/0.2) \cong 0.1510$. The estimated entropy of the utterance X_j reduces to $H(L_j|X_i) = 1.522(1-\mathrm{e}^{-0.1510}) \cong 0.213$ nats if the utterance X_i is selected for transcribing when α and β are set to 1.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

3. Semi-supervised learning with global entropy reduction maximization criterion

The key task in the semi-supervised learning is to choose the utterance-transcription pairs from the un-transcribed utterances so that the AM trained with these pseudo-transcriptions can achieve the best recognition accuracy. This task is usually simplified as selecting a best transcription from the lattice for an utterance, and determining whether the utterance-transcription pair would be beneficial in improving the AM. The existing algorithms typically use the top hypothesis as the pseudo-transcription and determines whether to trust (or use) the hypothesis based on the confidence score (e.g., posterior probability) of that hypothesis. This approach can work fine when the initial AM is of high quality but may fail when the recognition accuracy and the confidence score of the initial AM are poor.

We take a different perspective. We argue that the quality of the pseudo-transcription should be judged collectively with information contained in all the transcribed and un-transcribed utterances. Assume there are three acoustically similar utterances X_1, X_2, and X_3, and A and B are two possible pseudo-transcriptions for these utterances. The recognition results for X_1, and X_2, are $P_1(A) = 0.8$, $P_1(B) = 0.2$, $P_2(A) = 0.8$ and $P_2(B) = 0.2$. The recognition results for X_3 is $P_3(A) = 0.45$ and $P_3(B) = 0.55$. If we only depend on the confidence score of the single utterance, we would pick B as the pseudo-transcription of X_3 and use it in the training. However, if we also consider the other two utterances that are acoustically very close to X_3, we would more likely to choose A as the transcription for it or even do not use this utterance at all. Examine this condition more closely. We have two outcomes if A is chosen as the transcription of X_3. If A is the true transcription, adding it to the training set would increase its own confusability but decrease the confusability for the utterances X_1 and X_2. If B is the true transcription, using A as the transcription would decrease its own confusability but increase the confusability of the other two utterances. The average effect depends on the probabilities each condition would happen. This example suggests that we may measure how an utterance-transcription pair may affect the retrained system by measuring the expected entropy reduction the utterance-transcription pair can cause over the whole dataset.

Put formally, let $X_1, X_2, \ldots, X_n$ be the n candidate speech utterances. We wish to choose the best utterance-transcription pair $\{X_j, T_j\}$ that will have the maximum positive expected reduction of entropy in the lattices $L_1, L_2, \ldots, L_n$ over the whole dataset

$$E\left[\Delta H\left(L_1, \ldots, L_n | X_j, T_j\right)\right] \cong \sum_{i=1}^{N} E\left[\Delta H\left(L_i | X_j, T_j\right)\right] = \sum_{i=1}^{N} E\left[\Delta H^s_{i|j}\right], \tag{9}$$

where we have used the assumption that utterances are independently drawn. Note that similar to the active learning case, we need to adjust the current entropy after each selection.

To simplify the optimization problem, we have chosen to use the top hypothesis as the best possible transcription for each utterance at the current stage. We now describe how we may estimate $E[\Delta H^s_{i|j}]$ with pair-wise confusions between lattices by noting our key intuition: transcribing two acoustically similar utterances differently would increase the entropy.

Consider two utterances X_i and X_j. Let L_i and L_j be the recognition lattices obtained with the original AM Θ for these two utterances respectively. Let $\hat{L}_i$ be the transcription lattice obtained when decoding X_i with the AM trained using both the initial training set and the pair $\{X_j, T_j\}$ where T_j is a pseudo-transcription, which at the current stage is the best path in the lattice. We tabulate the pair-wise confusions present in these lattices by comparing the time-durations of every pair of nodes in the lattices. If the percentage overlap in the time duration is greater than a particular threshold, we say that the two nodes are getting confused. Note that the best path through the lattice is simply a sequence of words that give the highest likelihood. Out of these pair-wise confusions, we pick only those confusions which have a word/phone from the best path. Let $\{u_i^1, v_i^1\}, \{u_i^2, v_i^2\}, \ldots, \{u_i^{i_N}, v_i^{i_N}\}$ and $\{u_j^1, v_j^1\}, \{u_j^2, v_j^2\}, \ldots, \{u_j^{j_N}, v_j^{j_N}\}$ be the pair-wise confusions from the lattice of L_i and L_j, respectively, where u_i^k and v_i^k is a pair of arcs in the lattice L_i. u_i^k is an arc in the best path and v_i^k is the most confusing arc to u_i^k in the same lattice. Let $< \hat{b}_i^1, \hat{b}_i^2, \ldots, \hat{b}_i^{i_N} >$ and $< \hat{b}_j^1, \hat{b}_j^2, \ldots, \hat{b}_j^{j_N} >$ be the top hypothesis from the lattice L_i and L_j, respectively, where $\hat{b}_i^k$ is the kth word or phoneme in the top hypothesis, and $\{P(u_i^1), P(v_i^1)\}, \ldots, \{P(u_i^{i_N}), P(v_i^{i_N})\}$ and $\{P(u_j^1), P(v_j^1)\}, \ldots, \{P(u_j^{j_N}), P(v_j^{j_N})\}$ be the probabilities of these arcs on the lattices L_i and L_j based on the acoustic model score only, which we will use to compute the acoustic differences between two given signals.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

The pair-wise confusion can be computed at the word or phoneme level. In our experiments, we used the word lattices since the decoder we have used outputs word lattices. Given the fact that if $\{u_i^n, v_i^n\} = \{u_j^m, v_j^m\}$ and u_i is present in the best path of both the lattices L_i and L_j, then there will be an entropy reduction in L_i' which would be related to the distance between $\{P(u_i^n), P(v_i^n)\}$ and $\{P(u_j^m), P(v_j^m)\}$. If u_i is in the best path of L_i but v_i is in the best path of L_j, there will be a rise in entropy. We approximate the entropy reduction that $\{X_j, T_j\}$ would cause on L_i as

$$E\left[\Delta H_{i|j}^s\right] \cong -\alpha H_i \sum_{m=1}^{i_N} \sum_{n=1}^{j_N} \mathrm{e}^{-\beta d\left(\left\{P(u_i^m),P(v_i^m)\right\};\left\{P(u_j^n),P(v_j^n)\right\}\right)} (-1)^{I\left(\hat{b}_i^m=\hat{b}_j^n\right)} \tag{10}$$

where α and β are related to the training method used and the existing model, and may be estimated using the initial transcribed training set, and $d(\{P(u_i^m), P(v_i^m)\}; \{P(u_j^n), P(v_j^n)\}$ is the Kullback–Leibler divergence between the probability distributions $\{P(u_i^m), P(v_i^m)\}$ and $\{P(u_j^n), P(v_j^n)\}$. The net entropy change due to putting utterance X_j with its top hypothesis as the transcription into the training data is given as

$$E\left[\Delta H_j\right] = \sum_{i=1}^{N} E\left[\Delta H_{i|j}^s\right] \tag{11}$$

4. Unified procedure and framework

As we have illustrated in Sections 2 and 3, both the active learning and semi-supervised learning can be cast as a global entropy reduction maximization problem and can be carried out using the same procedure detailed in the following:

- Step 1: For each of the n candidate utterances, compute the entropy $H_1, H_2, \ldots, H_n$ from the lattice. If $\mathcal{Q}_i$ is the set of all paths in the lattice of the i^{th} utterance, the entropy can be computed as

$$H_i = -\sum_{q \in \mathcal{Q}_i} p_q \log(p_q) \tag{12}$$

  where p_q is the posterior probability of the path q in the lattice. This can be computed efficiently by doing a single backward pass. The entropy of the lattice is the entropy $H(S)$ of the start-node S. If $P(u, v)$ is the probability of going from node u to node v, the entropy of each node can be written as

$$H(u) = \sum_{v:P(u,v)>0} P(u,v)(H(v) - \log(P(u,v))) \tag{13}$$

  This simplifies the computation of entropy greatly where there are millions of paths and the computation is in $O(V)$ where V is the number of vertices in the graph.
- Step 2: If $H_1, H_2, \ldots, H_n$ are the entropy values for each of the n utterances, for each utterance X_i where $1 \leqslant i \leqslant n$, we compute the expected entropy reduction ΔH_i that this utterance will cause on all the other utterances using (6) for the active learning case, and (10) for the semi-supervised learning case, i.e.,

$$E[\Delta H_i] \cong \alpha \sum_{j=1}^{n} H_j \mathrm{e}^{-\beta d(X_i, X_j)} \tag{14}$$

  for the active learning case, and

$$E[\Delta H_i] \cong -\alpha \sum_{j=1}^{n} H_j \sum_{m=1}^{i_N} \sum_{n=1}^{j_N} \mathrm{e}^{-\beta d\left(\left\{P\left(u_i^m\right),P\left(v_i^m\right)\right\};\left\{P\left(u_j^n\right),P\left(v_j^n\right)\right\}\right)} (-1)^{I\left(\hat{b}_i^m=\hat{b}_j^n\right)} \tag{15}$$

  for the semi-supervised case.
- Step 3: Choose the utterance $X_{\hat{i}}$ which has not been chosen before and has the highest value of $E[\Delta H_i]$ among all the utterances.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

- Step 4: Update the values of the entropy after choosing $X_{\hat{i}}$ using

$$H_j^{t+1} \cong H_j^t - E\left[\Delta H_{j|i}\right]. \tag{16}$$

where $\Delta H_{j|i} = \Delta H^a_{j|i}$ for active learning, and $\Delta H_{j|i} = \Delta H^s_{j|i}$ for semi-supervised learning. Note that only the utterances that are close to $X_{\hat{i}}$ need to be updated. In this study, we have used the KLD as the distance and only updated the utterances X_j with $d(X_{\hat{i}}, X_j)$ less than or equal to 2.3. The threshold is so chosen that the change of the entropy is less than 10% of the original entropy.

- Step 5: Goto step 6 if k utterances have been chosen in the active learning case, or $E[\Delta H_i] < 0$ for all X_i in the semi-supervised case. Goto Step 2 otherwise.
- Step 6: (optional and is only for the active learning) The accuracy can be further improved if each selected utterance is weighted, for example by counting the utterances that are very close to it with the distance we have already defined. A heuristic we have used is to use

$$w_i \propto \sum_{j \in R(i)} \mathrm{e}^{-\beta d\left(X_i, X_j\right)}, \tag{17}$$

where $R(i)$ is the set of utterances that have not been selected for transcribing and are closer to X_i than to all other utterances selected.

5. Experimental results

We have evaluated our algorithm using the directory assistance data collected under the real usage scenarios. The 39-dimentional features used in the experiments were converted with HLDA from a 52-dimensional feature – a concatenation of the 13-dimension MFCC, its first, second, and third derivatives. We did not tune α and β in these experiments and simply set them to one. The initial AM was trained with maximum likelihood (ML) criterion using around 4000 utterances and was used to generate the lattices for the candidate utterances, the candidate set consists of around 10,000 and 30,000 utterances for two different settings, and the test set contains around 10,000 utterances. We have tested with other settings with more or less data and got similar improvements.

5.1. Active learning

To compare our new active learning algorithm with existing confidence-based algorithms, we selected 1%, 2%, 5%, 10%, 20%, 40%, 60%, and 80% of the candidate utterances using the active learning algorithms, combined them with the initial training set, and retrained the model with ML criterion. We have used two baselines in the experiments: the random sampling approach and the confidence-based approach. The random sampling approach selects the top k utterances randomly. We ran the random sampling 10 times and report the mean of the 10 runs. The standard deviation of the 10 runs is between 0.01% and 0.07% depending on the percentage selected with an average standard deviation of 0.03%. The confidence-based approach selects the least confident k utterances for transcribing. There can be many ways to computing confidence scores (Riccardi and Hakkani-Tur, 2005, e.g.,; Zhang and Rudnicky, 2001). In our experiments we have used the lattice entropy and the posterior probability as the confidence and achieved similar results.

We have evaluated the GERM algorithm proposed in this paper both with and without the weighing described in the step 6 of the unified procedure. Fig. 1 compares the GERM algorithm with the random sampling approach and the confidence-based approach using the 10,000 candidate set. From Fig. 1, we can see that the GERM algorithm with and without the weighting both consistently outperform the confidence-based approach with a significant margin. Under the condition where a fixed amount of data are allowed to be transcribed, our approach without the weighting outperforms the confidence-based approach by maximum of 2.3% relatively. To achieve the same accuracy, our approaches can cut the number of utterances needed for transcribing by 50% compared to the confidence-based approach and by 60% compared to the random sampling approach. All these improvements are statistically significant at significance level of 1%. From Fig. 1 we can also see that the GERM algorithm with weighting slightly outperforms the approach without the weighting.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

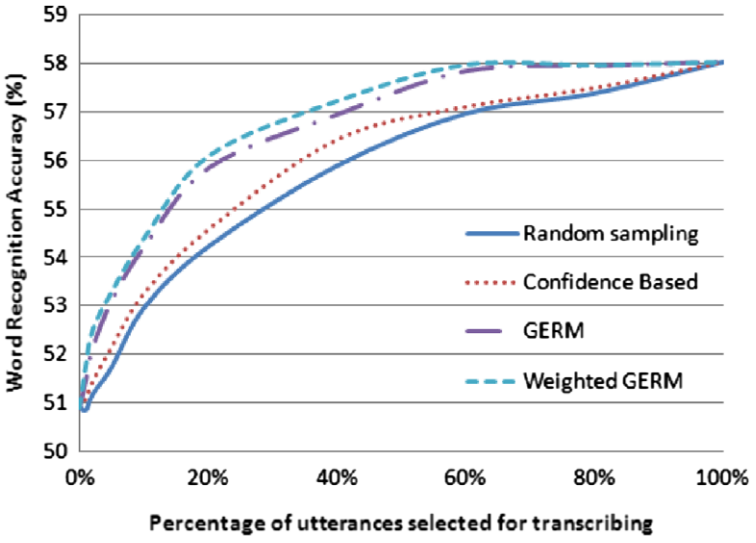

Fig. 1. Speech recognition accuracies (%) among different active learning approaches with the 10,000 utterances candidate set when different percentage of utterances are allowed to be transcribed.

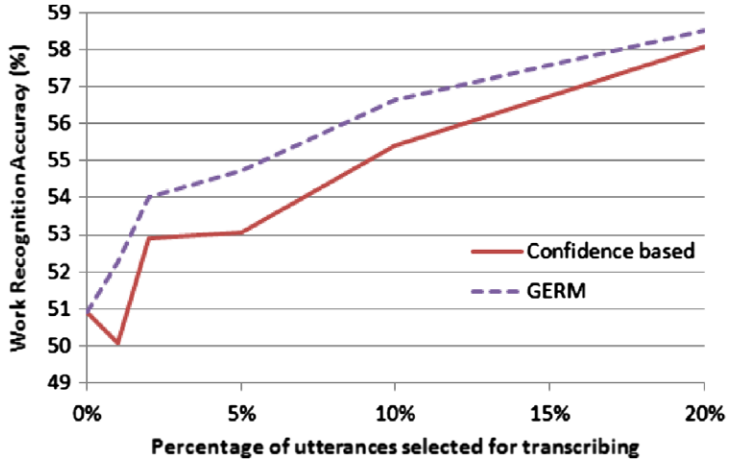

Fig. 2. Speech recognition accuracies (%) between confidence-based active learning approach and GERM-based approach with the 30,000 utterances candidate set when different percentage of utterances are allowed to be transcribed.

To better understand the algorithm, we have manually checked the utterances selected by the confidence-based approach and the GERM algorithm. We observed that if only 1% of utterances are to be selected, most utterances selected by the confidence-base approach are noise and garbage utterances that have extremely low confidence but have little value to improving the performance of the overall system, while only a few such utterances are selected by the GERM algorithm. This difference is demonstrated in Fig. 2 where we have used 30,000 utterances as the candidate set. Note that, the performance becomes worse if 1% of the un-transcribed data selected by the traditional confidence-based approach are transcribed. This is not the case if the GERM algorithm was used. This observation further confirmed the superiority of the GERM algorithm.

5.2. Semi-supervised training

We have also conducted experiments to see how good the criterion we are using in the semi-supervised training is compared to that in confidence-based approaches. To do this, we used the initial AM to generate

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

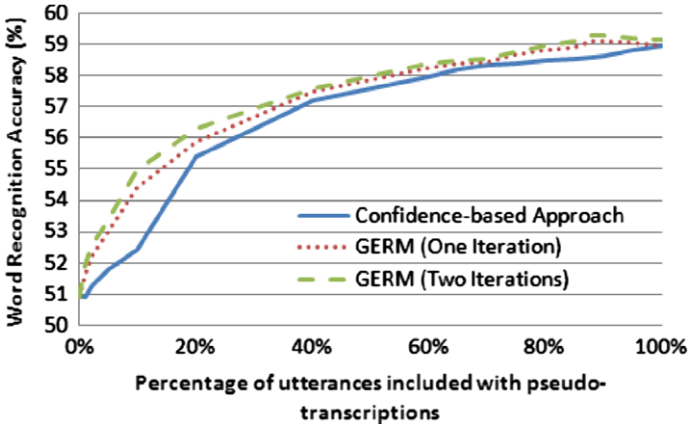

Fig. 3. Compare speech recognition accuracies (%) between different semi-supervised learning approaches with the 30,000 utterances candidate set when top k% of utterance-transcription pairs are used in the training.

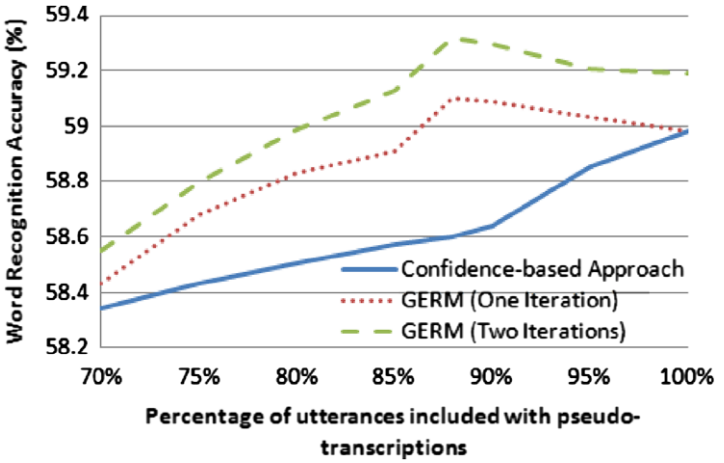

Fig. 4. Speech recognition accuracies (%) between different semi-supervised learning approaches with the 30,000 utterances candidate set focusing on the peak area.

the lattices for the un-transcribed utterances. We then selected 1%, 2%, 5%, 10%, 20%, 40%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, and 100% from the 30,000 candidate utterances using different semi-supervised learning algorithms, combined them with the initial training set, and retrained the model with ML criterion. The dotted red curve and the solid blue curve in Figs. 3 and 4 compare the results obtained with our algorithm and that with the traditional confidence-based approach.

There are three important observations in this comparison. First, there is no peak using the confidence-based approach. Adding new utterances continues to improve the recognition accuracy. Using our newly developed algorithm, however, we do observe a peak around 86% position (which is easier to be noticed in Fig. 4). This indicates that the ranking from our algorithm is better than that from the confidence-based approach. In other words, our algorithm has better ability to find good pseudo-transcriptions and rule out bad ones than the confidence-based approach. Note however, although there is a peak using our approach, the peak is not very far away from 100%. This is due to the fact that the accuracy of the initial AM is very low and so the posterior probabilities in the lattices are also very poor.

Second, not only is there a peak using the GERM based semi-supervised algorithm, but also the peak can be estimated. As we have discussed in Section 3, a negative expected entropy reduction indicates that adding the utterance might make the recognizer worse. The cutoff point found by this principled threshold is 88% on

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

this task and the corresponding accuracy number is 59.1%. The cutoff point found is very close to the true peak point shown in the figures. The threshold found is task dependent. However, the approach can be generalized to other tasks.

Third, we can observe that if the same amount of utterances is selected, our algorithm consistently outperforms the confidence-based approach and the differences are statistically significant at the significance level of 1%. This is another indication that the criterion and algorithm proposed in this paper is superior to the confidence-based approach. Note that we have not yet investigated the use of the hypothesis other than the top one and did not tune any of the parameters used in the algorithm. We believe better results can be achieved once we integrate all these into the algorithm.

Our algorithm can be integrated into either the incremental training or GEM training strategy. To see what performance we may get with the incremental training, we have retrained the AM with 88% (which is the value automatically determined by our algorithm) of the pseudo-transcriptions, regenerated the lattices for all the candidate utterances, determined and selected the new pseudo-transcriptions, and retrained the AM. We achieved 59.32% accuracy, which is 0.2% better than the first iteration. If we train the AM with 100% *true* transcriptions, we can get the upper bound which is 61.06%. The dotted red curve and the dashed green curve in Fig. 3 compares the results using our proposed approach with one and two iterations. It can be seen that the second iteration is slightly better than the first iteration because a better acoustic model (the result of the first iteration) was used in the second iteration.

5.3. Combine active learning and semi-supervised learning

In our last set of experiments we combined the active and semi-supervised learning with three different settings. In setting 1, we first use our active learning algorithm to select x% of the data for supervised training and use the semi-supervised training algorithm to select the pseudo-transcription for the remaining 100-x% utterances, all with the initial AM. In the setting 2, we retrain the AM after the active learning step, decode the remaining 100-x%, then use our semi-supervised learning algorithm to select the pseudo-transcriptions for the remaining 100-x% utterances. In the setting 3, we did the same as in the setting 2 but ran the semi-supervised learning algorithm for two iterations. Fig. 5 illustrates the result we have obtained with the 30,000 utterances candidate set. There are three observations. First, by combining two approaches we can obtain 60.15% recognition accuracy by transcribing only 20% of the data. This is especially good considering that the best we can get is 61.06% with all data transcribed. Second, retraining the AM after the active learning step helps most when x is in the mid-range ($[10, 70]$ in this case). We believe that this is because when x is small (less than 10 in this case), retraining does not change the AM too much and so won't greatly affect the pseudo-transcription

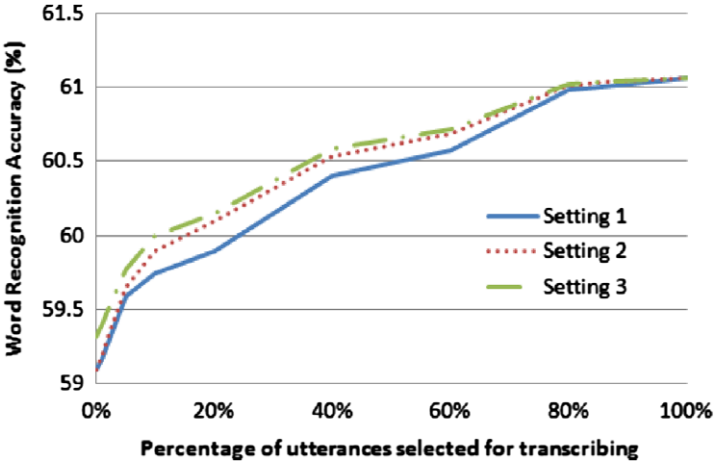

Fig. 5. Speech recognition accuracies (%) under different settings when our new active learning and semi-supervised learning algorithms are combined (tested on the 30,000 utterances candidate set). The x-axis shows the percentage of utterances selected by the active learning algorithm.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

obtained in the semi-supervised learning step. When x is large (greater than 70 in this case), on the other hand, the number of utterances left for semi-supervised learning is small and so the slight difference in the pseudo-transcription would not greatly affect the resulting AM either. Third, running the semi-supervised learning for two iterations helped much when x is small. This is because as x becomes larger, the AM retrained after the active learning step has closer performance as the AM trained after the first iteration of the semi-supervised learning.

6. Summary and conclusion

We have described a unified framework for active learning and semi-supervised learning for speech recognition. The core idea of our framework is to select the utterances in the active learning case or the utterance-transcription pairs in the semi-supervised case, so that the uncertainties for the whole dataset can be minimized. This global entropy reduction maximization based framework can be justified by the fact that a better decision can be made if information from all the utterances are taken into account. We showed the simplifications and approximations made to make the problem tractable. The effectiveness of our algorithm was demonstrated using the directory assistance data recorded under the real usage scenarios. The experiments indicated that our new active learning algorithm can cut the number of utterances by 50% to achieve the same accuracy obtained with the confidence-based approach, and by 60% compared with the random sampling approach. The experiments also demonstrate that our new semi-supervised learning algorithm has better ability to identify the good utterance-transcription pairs than the confidence-based approaches and can automatically identify the cutoff point. By combining active learning and semi-supervised learning algorithms, we can achieve even better results.

There are many areas to improve along this line of research. For example, we have not utilized any hypothesis other than the top one in our current semi-supervised algorithm and experiments, and the approximation we have made is rather crude. We will further improve the system in the future work.

References

Anderson, B., Moore, A., 2005. Active learning for hidden markov models: objective functions and algorithms. In: ICML 2005.

Anderson, B., Siddiqqi, S., Moore, A., 2006. Sequence Selection for Active Learning. Technical Report CMU-IR-TR-06-16.

Charlet, D., 2001. Confidence-measure-driven unsupervised incremental adaptation for HMM-based speech recognition. In: Proceedings of ICASSP. pp. 357360.

Cohn, D., Atlas, L., Ladner, R., 1994. Improving generalization with active learning, machine learning 15 (2), 201221.

Dagan, I., Engelson, S.P., 1995. Committee-based sampling for training probabilistic classifiers. In: Proceedings of ICML. pp. 150157.

Hakkani-Tr, D., Tur, G., Rahim, M., Riccardi, G., 2004. Unsupervised and active learning in automatic speech recognition for call classification. In: ICASSP 2004.

Hakkani-Tur, D., Gorin, A., 2002. Active learning for automatic speech recognition. In: Proceedings of the ICASSP. pp. 39043907.

Ji, S., Krishnapuram, B., Carin, L., 2006. Variational bayes for continuous hidden markov models and its application to active learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4), 522–532.

Kamm, T.M., 2004. Active Learning for Acoustic Speech Recognition Modeling, Ph.D. Thesis. The Johns Hopkins University, Baltimore.

Kamm, T.M., Meyer, G.G.L., 2003. Word-selective training for speech recognition. In: Proceedings of the IEEE Workshop ASRU.

Kamm, T.M., Meyer, G.G.L., 2004. Robustness aspects of active learning for acoustic modeling. In: Proceedings of Interspeech, pp. 1095–1098.

Kemp, T., Waibel, A., 1999. Unsupervised training of a speech recognizer: recent experiments. In: Proceedings of Eurospeech. pp. 27252728.

Kuo, H.-K.J., Goel, V., 2005. Active learning with minimum expected error for spoken language understanding. In: Proceedings of the Interspeech. pp. 437440.

Moreno, P.J., Agarwal, S., 2003. An experimental study of em based algorithms for semi-supervised learning in audio classification. In: ICML-2003 Workshop on Continuum from Transcribed to Un-transcribed Data.

Riccardi, G., Hakkani-Tur, D., 2005. Active learning: theory and applications to automatic speech recognition. IEEE Transactions on Speech and Audio Processing 13 (4), 504511.

Tur, G., Hakkani-Tr, D., Schapire, R.E., 2005. Combining active and semi-supervised learning for spoken language understanding. Journal of Speech Communication 45 (2), 171–186.

Wessel, F., Macherey, K., Ralf Schluter, 1998. Unsupervised training of a speech recognizer: recent experiments. In: Proceedings of ICASSP. pp. 225228.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004

Yu, D., Ju, Y.-C., Wang, Y.-Y., Zweig, G., Acero, A., 2007. Automated directory assistance system – from theory to practice. In: Proceedings of the Interspeech. pp. 27092712.

Zhang, R., Rudnicky, A., 2001. Word level confidence annotation using combinations of features. In: Proceedings of ECSCT. pp. 21052108.

Zhang, R., Rudnicky, A.I., 2006. A new data selection approach for semi-supervised acoustic modeling. In: ICASSP. pp. 421424.

Please cite this article in press as: Yu, D. et al., Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion, Computer Speech and Language (2009), doi:10.1016/j.csl.2009.03.004