MAXIMIZING GLOBAL ENTROPY REDUCTION FOR ACTIVE LEARNING IN SPEECH RECOGNITION

Balakrishnan Varadarajan[*]

Johns Hopkins University
3400 North Charles Street
Baltimore, MD 21218
bvarada2@jhu.edu

Dong Yu, Li Deng, Alex Acero

Microsoft Research
One Microsoft Way
Redmond, WA 98052
{dongyu, deng, alexac}@microsoft.com

ABSTRACT

We propose a new active learning algorithm to address the problem of selecting a limited subset of utterances for transcribing from a large amount of unlabeled utterances so that the accuracy of the automatic speech recognition system can be maximized. Our algorithm differentiates itself from earlier work in that it uses a criterion that maximizes the lattice entropy reduction over the whole dataset. We introduce our criterion, show how it can be simplified and approximated, and describe the detailed algorithm to optimize the criterion. We demonstrate the effectiveness of our new algorithm with directory assistance data collected under the real usage scenarios and show that our new algorithm consistently outperforms the confidence based approach by a significant margin. Using the 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.

Index Terms— Active learning, acoustic model, entropy, confidence, lattice

1. INTRODUCTION

With the increased deployment of interactive voice response (IVR) systems (e.g., voice search applications[1]) collecting a large amount of unlabeled speech data becomes as easy as logging the interaction in a database. Transcribing these data for supervised training, however, is usually costly. For example, it may take a transcriber one month to transcribe one day of speech data. Optimally determining the subset for transcribing is thus very important to further improve the performance of the deployed systems.

This data selection problem is often casted as an active learning problem, where a question is actively asked so that some criterion can be optimized when the answer to the question becomes known. Specific to the data selection problem we tackle in this paper, we want to determine which subset of k utterances $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ should be selected from a total of n utterances $\{x_1, x_2, \ldots, x_n\}$ so that we may maximize the recognition accuracy with the retrained acoustic model (AM) on the unseen test set when the transcriptions of the selected utterances become known.

Active learning has been studied for decades and many approaches have been proposed. The approaches that have been successfully used in spoken dialog systems [2] and automatic speech recognition (ASR) systems [3] [4] can be classified into three categories: confidence-based approach [3] [4], query-by-committee approach [5], and error-rate reduction approach [2]. In the confidence-based approach, utterances with the least confidence are selected for transcribing. In the query-by-committee approach, utterances that cause biggest different opinions from a set of recognizers (committee) are selected, and in the error-rate reduction approach, the utterances that can minimize the expected error rate most are selected.

In this paper we propose a new and improved active learning algorithm for speech recognition. The algorithm falls into the category of confidence-based approaches. However, different from the existing confidence-based approaches, our algorithm, which is named as global entropy reduction maximization (GERM) algorithm, uses a criterion that maximizes the lattice entropy reduction over the whole dataset. More specifically, the GERM algorithm measures the Kullback–Leibler divergence (KLD) between lattices generated by decoding the unlabeled utterances, estimates the expected entropy reduction over the whole dataset for each given utterance, and selects the utterances that can cause the highest entropy reduction over the whole dataset for transcribing. Furthermore, the transcribed utterances can be weighted according to the number of similar utterances in the whole dataset to achieve better performance. We evaluated our algorithm using the directory assistance [1] data collected under the real usage scenarios. Our experiments show that the GERM algorithm outperforms the traditional confidence-based approach by a significant margin over all settings and

[*]This work was carried out during the internship program at Microsoft research.

can cut the number of utterances needed for transcribing by 50% to achieve the same recognition accuracy obtained using the earlier confidence-based approach, and by 60% compared with the random sampling approach.

The rest of the paper is organized as follows. In Section 2 we discuss the limitations of the existing confidence-based approaches and introduce the new criterion used in our algorithm. In Section 3 we describe the GERM algorithm in detail, with the focus on the simplifications and approximations used. We present our experimental results in Section 4 and conclude the paper in Section 5.

2. THE GERM CRITERION

As has been pointed out in Section 1, the existing confidence-based approaches select the utterances that are least confident for transcribing. They are based on the heuristics that transcribing the least confident ones can provide the most information to the system.

While selecting the least confident utterances seems to be reasonable at the first glance, limitations can be observed under careful examination esp. when applied to the spontaneous speech utterances recorded under real usage environments. For example, we have observed a large collection of noise and garbage utterances in the directory assistance dataset. These utterances typically have low confidence scores and will be selected for transcribing by the confidence-based approach. However, transcribing these utterances is usually difficult and carries little value in improving the ASR performance.

The above limitation of the existing confidence-based approaches comes from the fact that the information from a selected utterance may not be useful to improve the performance of other utterances. Consider two speech utterances A and B. A has a slightly lower confidence score than B has. However if A is observed only once and B occurs frequently in the dataset transcribing B would correct a larger fraction of errors in the test data than transcribing A and thus has higher probability to improve the performance of the whole system. A reasonable choice is thus to transcribe B instead of A as will be selected by the confidence-based approaches. This example brings up the notion that we should select the utterances that can achieve most for the whole dataset and this is the core idea of our new algorithm.

Using a global criterion has been explored by Kuo and Goel [2] for the dialog system upon the error-rate reduction approaches. The GERM algorithm proposed in this paper differs from their approach in that we use a different criterion that would maximize the expected lattice entropy reduction over all the unlabeled data from which we wish to select. Optimizing the entropy is more robust than optimizing the top choice since it considers all possible outputs weighted with probabilities. In addition, ASR is a sequential recognition problem where we need to consider the segments in the lattices or recognition results when estimating the gains and thus is a much more difficult scenario than the static classification problem Kuo and Goel focused on.

Now let us define our active learning criterion formally. Let $X_1, X_2, \ldots, X_n$ be the n candidate speech utterances. 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

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

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

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

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

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. Since $H(L_1, \ldots, L_n|\Theta)$ is a fixed value, maximizing (4) is equivalent to minimizing the expected entropy under the new model

$$E[H(L_1, \ldots, L_n|\Theta^s)] \quad (5)$$

Note that this optimization problem is NP-hard 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 similar to the chosen one is changed significantly. We approximate the solution to this optimization problem with a greedy algorithm with which we select a single utterance that maximizes the expected entropy reduction over the whole dataset. We then adjust the entropies for all similar utterances and determine the next utterance that gives us the highest gain, and so on.

3. ALGORITHM DESCRIPTION

3.1. Simplifications

The key formula to evaluate in our approach is the expected entropy reduction (4) when an utterance X_i is selected for transcribing, which we will approximate using a distance based approach using the following two assumptions.

First, we assume that the expected entropy reduction on L_i is proportional to its original entropy, or

$$E[\Delta H(L_i|X_i)] \cong \alpha H(L_i|\theta), \quad (6)$$

where α is a parameter related to the training algorithm used and the number of transcribed utterances in the initial training set and may be estimated from the training set.

Second, we assume that the impact of utterance X_i on utterance X_j is a function of the distance $d(X_i, X_j)$ between utterances X_i and X_j. In the extreme case, if the utterance X_i and its transcription T_i is given and the transcription T_i does not contain any phone that is present in lattice L_j, the AM of

any of the phones in the lattice L_j will not be updated. This implies that the acoustic scores and hence the probabilities of all the paths in the lattice L_j will remain the same, or

$$E[\Delta H(L_j|X_i)] = 0. \tag{7}$$

In a more general case, we approximate the expected entropy reduction over L_j with X_i selected for transcribing as

$$E[\Delta H(L_j|X_i)] \cong \alpha H(L_j|\Theta)e^{-\beta d(X_i,X_j)} \tag{8}$$

where α and β can be estimated from the initial transcribed training set, $d(X_i, X_j) = 0$ if two utterances are the same and $d(X_i, X_j) = \infty$ if two utterances do not have common phones in the lattices. This distance $d(X_i, X_j)$ can be estimated in several ways including the dynamic time warping (DTW) distance between the utterances X_i and X_j. In this paper we have used the KLD between two lattices of L_i and L_j as the distance. For example if lattices L_i and L_j both confuse between words star, stark and start with probabilities $P_i(star) = 0.4$, $P_i(stark) = 0.2$, $P_i(start) = 0.2$ and $P_j(star) = 0.3$, $P_j(stark) = 0.3$, $P_j(start) = 0.4$. The initial entropy of lattice L_j is 0.473 nats. The distance between two lattices is estimated as $d(X_i, X_j) = KLD(0.3, 0.3, 0.4; 0.4, 0.2, 0.2) \approx 0.1375$. The estimated entropy of the utterance X_j reduces to $H(L_j|X_i) = 0.473(1 - e^{-0.1375}) \cong 0.06$ if the utterance X_i is selected for transcribing when α and β are set to 1.

Given (8), the expected entropy reduction over the whole dataset can be approximated as

$$E[\Delta H(L_1, \ldots, L_n|X_i)] \cong \tag{9}$$

$$\sum_{j=1}^{n} E[\Delta H(L_j|X_i)] \cong \tag{10}$$

$$\alpha \sum_{j=1}^{n} H(L_j|\Theta)e^{-\beta d(X_i,X_j)} \tag{11}$$

where we have assumed that the utterances are independently drawn. Our objective now becomes to choose an utterance X_i maximizing (11) at each step, update the expected entropies after the X_i is chosen, and then select the next best utterance based on (11) with the updated entropies.

3.2. Procedure

Our algorithm can be summarized in the following steps:

- 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

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

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

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

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 $1 \leq i \leq n$, we compute the expected entropy reduction ΔH_i that this utterance will cause on all the other utterances using (11), i.e.,

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

- Step 3: Choose the utterance $X_{\hat{i}}$ which has not been chosen before and has the highest value of ΔH_i among all the utterances.

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

$$H_i^{t+1} \cong H_i^t \left(1 - \alpha e^{-\beta d(X_i,X_{\hat{i}})}\right). \tag{15}$$

Note that only the utterances that are close to $X_{\hat{i}}$ need to be updated.

- Step 5: Goto step 6 if k utterances has been chosen, otherwise goto Step 1.

- Step 6: (optional) 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

$$w_i \propto \sum_{j \in R(i)} e^{-\beta d(X_i,X_j)}, \tag{16}$$

where $j \in R(i)$ if and only if j is not selected for transcribing and is closer to X_i than to all other utterances selected.

4. EXPERIMENTAL RESULTS

We have evaluated our algorithm using the directory assistance data, which are spontaneous speech collected under various background noises and channel distortions. The vocabulary size is 100K. The 39-dimentional features used in the experiments were converted with HLDA from a 52-dimensional

feature concatenated with 13-dimention MFCC, its first, second, and third derivatives. In the results reported in Figure 1, the initial AM was trained with maximum likelihood (ML) using around 4000 utterances, the candidate set consists of around 10000 utterances, and the test set contains around 10000 utterances. We have tested with other settings with more data and got the similar improvements.

The initial model was used to generate the lattices for the candidate utterances. We then 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. Two baselines were used 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

We have evaluated the GERM algorithm proposed in this paper both with and without the weighing. We didnt tune the α and β in these experiments and simply set them to 1. Figure 1 compares the GERM algorithm with the random sampling approach and confidence-based approach. From Figure 1, we can see that the GERM algorithm with weighting slightly outperforms the approach without the weighting, and both outperform the confidence-base approach with a significant margin consistently. For the same amount of data selected for transcribing, our approaches outperform 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%.

To better understand the algorithm, we have manually checked the utterances selected by the confidence-based approach and the GERM algorithm. We have 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 improve the performance of the overall system, while only a few such utterances are selected by the GERM algorithm. This observation further confirmed the superiority of the GERM algorithm.

5. SUMMARY AND CONCLUSIONS

We have described a new active learning algorithm for improving acoustic models. The core idea of our algorithm is to select the utterances that have the highest impact in reducing the uncertainties for the whole dataset. 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 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.

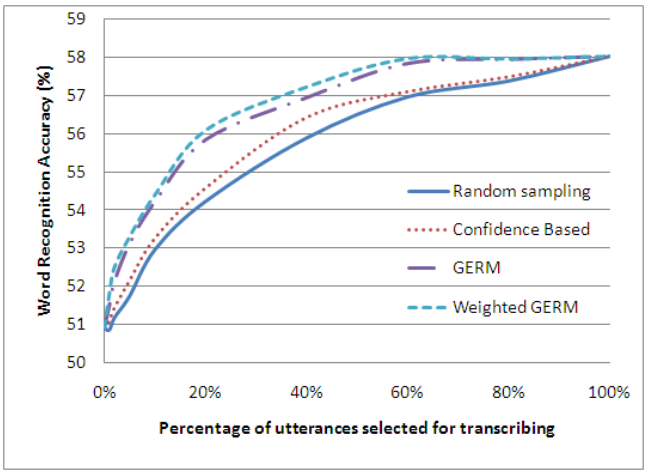

Fig. 1. Compare speech recognition accuracies among different approaches

6. ACKOWLEDGEMENT

We owe special thanks to Dr. Patrick Nguyen and Geoffrey Zweig from Microsoft Speech research group for their technical help. We also like to thank Dr. Jasha Droppo for his help in handling the computing resources which made us do these experiments.

7. REFERENCES

[1] D. Yu, Y.-C. Ju, Y.-Y. Wang, G. Zweig, and A. Acero, "Automated directory assistance system - from theory to practice," in *Proc. of Interspeech*, 2007, pp. 2709–2712.

[2] H.-K. J. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding," in *Proc. of Interspeech*, 2005, pp. 437–440.

[3] D. Hakkani-Tur and A. Gorin, "Active learning for automatic speech recognition," in *Proc. of ICASSP*, 2002, pp. 3904–3907.

[4] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.

[5] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proc. of ICML*. 1995, pp. 150–157, Morgan Kaufmann.