USING COLLECTIVE INFORMATION IN SEMI-SUPERVISED LEARNING FOR 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

Training accurate acoustic models typically requires a large amount of transcribed data, which can be expensive to obtain. In this paper, we describe a novel semi-supervised learning algorithm for automatic speech recognition. The algorithm determines whether a hypothesized transcription should be used in the training by taking into consideration collective information from all utterances available instead of solely based on the confidence from that utterance itself. It estimates the expected entropy reduction each utterance and transcription pair may cause to the whole unlabeled dataset and choose the ones with the positive gains. We compare our algorithm with existing confidence-based semi-supervised learning algorithm and show that the former can consistently outperform the latter when the same amount of utterances is selected into the training set. We also indicate that our algorithm may determine the cutoff-point in a principled way by demonstrating that the point it finds is very close to the achievable peak point.

Index Terms— Semi-supervised learning, entropy reduction, lattice, confidence, collective information

1. INTRODUCTION

Training automatic speech recognition (ASR) systems usually requires a large amount of domain specific transcribed training data. However, getting transcribed data is usually very expensive and can be time consuming. On the other hand, getting un-transcribed data can be as easy as logging the data into a database. The goal of semi-supervised learning is to use both the transcribed and un-transcribed data to boost the performance of the ASR systems.

The application of the semi-supervised learning techniques to the ASR is not new [1, 2, 3, 4, 5]. Typical approaches used in the past include incremental training and the generalized expectation maximization (GEM). In the incremental training, the highly confident utterances are combined with transcribed utterances to adapt or retrain the recognizer and then use the adapted recognizer to select the next batch of utterances. In the GEM all utterances are used but with different weights determined by the confidence. Note that both these methods are confidence based. They either select or highly weight the utterances that have high confidence scores. Thus the sentences that are more likely to be correct are chosen to be in the training data. 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 [5] if the performance of the current model is very poor which is typically the case for the real-world interactive voice response systems (IVR) such as voice search applications [6].

In this paper, we propose a novel semi-supervised learning algorithm. Different from the previous approaches, our algorithm determines whether a hypothesized transcription should be used in the training by considering the collective information from all utterances available as opposed to only looking at its confidence. It estimates the expected entropy reduction on the unlabelled data set when a particular utterance and its transcription pair are chosen. We are particularly interested in scenarios where the amount of the transcribed data is small yielding a poor acoustic model. This is usually the case when a new application is developed. We compare our algorithm with confidence-based semi-supervised learning algorithm using the directory assistance data collected under the real usage scenarios. We show that our algorithm can consistently outperform the existing algorithm when the same amount of utterances is selected into the training set. We also indicate that our algorithm may determine the cutoff-point in a principled way by demonstrating that the point it finds is very close to the achievable peak point.

The rest of the paper is organized as follows. In Section 2, we introduce the criterion used in our new algorithm and describe the algorithm in detail. In Section 3 we compare our algorithm with the confidence-based approach empirically using the directory assistance data. We conclude the paper in Section 4.

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

2. SEMI-SUPERVISED ALGORITHM USING COLLECTIVE INFORMATION

2.1. Intuition and Criterion

The key problem in semi-supervised learning is to determine the utterance and transcription pair so that the acoustic model (AM) trained with these data can be optimized. Since we have some transcribed data to start with we can train an AM and use the generated recognizer to hypothesize the most possible transcriptions that are usually represented as a lattice. The problem of determining the transcription can thus be reduced to selecting a best transcription from the lattice. The existing algorithms typically use the top hypothesis and determines whether to trust (or use) the hypothesis based on the confidence score (e.g., posterior probability) of the hypothesis. This is typically fine if the initial AM is of high quality but is not a good solution when the recognition accuracy and thus the confidence score of the initial AM are poor.

In this paper, we take a different perspective. We argue that the quality of the hypothesis should be determined collectively by all the transcribed and un-transcribed utterances available. Assume three utterances X_1, X_2, and X_3 are very similar acoustically. 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 and assume the threshold is 0.5, we would pick B as the transcription of the utterance 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. This example suggests that we may measure whether a hypothesized transcription is appropriate by measuring how it will affect itself and other utterances.

Since we have the approximated probability of each condition from the recognizer, we may estimate the expected entropy reduction over the whole dataset for each possible hypothesis and use it as the measurement for the hypothesis. Note that we should not use a hypothesis if it will cause a negative expected entropy reduction.

Put it formally. Let $X_1, X_2, \ldots, X_n$ be the n candidate speech utterances. We wish to choose the best transcription T_j for each utterance X_j such that each selected utterance along with its suggested transcription will have the maximum positive expected reduction of entropy in the lattices $L_1, L_2, \ldots, L_n$ over the whole dataset

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

where we have used the assumption that utterances are independently drawn. Note that transcription T_j selected for utterance X_j may be right or wrong and that is the reason we optimize the expected (averaged) value of the entropy reduction.

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. The key formula to evaluate in our approach is the expected entropy reduction when a transcription is chosen for an utterance, which we will approximate as a function of the distance between the utterances.

We have already stated our key intuition: If two utterances that are acoustically similar are transcribed differently, that would result in increasing 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 hypothesized transcription, which in the current stage is the best path in the lattice. Now we can tabulate the confusions that are present in these lattices.

For simplicity, we tabulate the pair-wise confusions present in these lattices. This is obtained 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 at the current stage. Let $\{u_i^1, v_i^1\}$,$\{u_i^2, v_i^2\}$,...,$\{u_i^{i_N}, v_i^{i_N}\}$ and $\{u_j^1, v_j^1\}$,$\{u_j^2, v_j^2\}$,...,$\{u_j^{j_N}, v_j^{j_N}\}$ be the pair-wise confusions from the lattice of L_i and L_j respectively. Let $\{\hat{b}_i^1$,$\hat{b}_i^2$,...,$\hat{b}_i^{i_N}\}$ and $\{\hat{b}_j^1$,$\hat{b}_j^2$,...,$\hat{b}_j^{j_N}\}$ be the top hypothesis from the lattice L_i and L_j respectively, and $\{P(u_i^1), P(v_i^1)\}$,...,$\{P(u_i^{i_N}), P(v_i^{i_N})\}$ and $\{P(u_j^1), P(v_j^1)\}$,...,$\{P(u_j^{j_N}), P(v_j^{j_N})\}$ 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 difference between two given signals.

The units in the pair-wise confusion can be words or phones. 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

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

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 $\{P(u_j^n), P(v_j^n)\}$. Note that we have used the exponential function to approximate the true effect so that the effect is 1 when the KL-divergence is 0 and the effect is close to 0 when the KL-divergence is very large. The net entropy change due to putting utterance X_j with its top hypothesis as the transcription into the training data is given by

$$E[\Delta H_j] \cong \sum_{i=1}^{N} E[\Delta H_{i|j}] \quad (1)$$

2.2. Procedure

The algorithm proceeds as follows:

- Step 1: For each of the n candidate utterances, compute entropies $H_1, H_2, \ldots, H_n$ 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

  $$H_i \cong -\sum_{q \in \mathcal{Q}_i} p_q \log(p_q) \quad (2)$$

  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

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

  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: For each of the candidate utterances ($1 \leq j \leq N$), compute $E[\Delta H_j]$ as in (1).

- Step 3: Pick the utterance ($X_{\hat{j}}$) that has the maximum positive value among $E[\Delta H_j]$.

- Step 4: Update entropies for the utterances that are close to $X_{\hat{j}}$ using

  $$H_i^{t+1} \cong H_i^t - E[\Delta H_{i|\hat{j}}] \quad (3)$$

- Step 5: Stop if all the utterances are picked or $E[\Delta H_j] < 0$ for all j, otherwise goto Step 3.

3. EXPERIMENTAL RESULTS

We have evaluated our algorithm using the directory assistance data, which are spontaneous speech collected under various background noises and channel distortions [6]. The vocabulary size is about 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. The initial AM was trained with maximum likelihood (ML) using around 4000 utterances, the candidate set consists of around 30000 utterances, and the test set contains around 10000 utterances.

We compared our algorithm wih the confidence-based approaches, which augment the training set with the most confident utterances. In this experiment we have used the posterior probability as the confidence. We have also tried other confidence measures and achieved similar results. For example we have observed that the rankings based on the entropy of the lattice and that obtained based on the posterior probability of the most probable path from the lattice are highly correlated with a Spearman correlation coefficient between the two rankings greater than 0.92.In all the experiments we have conducted, we did not tune the α and β and simply set them to one.

Our first experiment is to see how well the criterion we are using is compared with the confidence-based approaches. In other words, can our criterion do better in selecting the utterances if the same amount of the utterances is selected? To do this, we used the initial AM to generate 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% of the candidate utterances using different semi-supervised learning algorithms, combined them with the initial training set, and retrained the model with the ML criterion. The dotted red curve and the solid blue curve in Figure 1 compares the results obtained with our algorithm and those 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. We believe this is due to the fact that the accuracy of the initial AM is very low and so the quality of the top hypothesis and the confidence score is also poor. In other words, the confidence score does not serve a good indicator to determine which utterance should be selected, and within each new batch of the data selected, the benefit from the partially right hypothesis always outweighs the bad effects. Using our newly developed algorithm, however, we do observe a peak around 90%. This indicates that our algorithm has better ability to rule out bad utterances and transcriptions than the confidence-based approach.

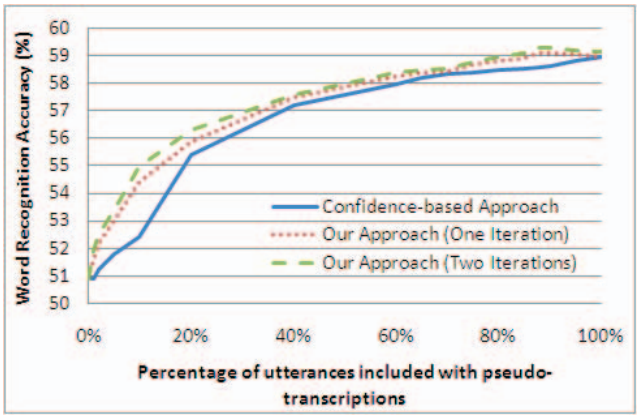

Fig. 1. Compare speech recognition accuracies between our new algorithm and the confidence based approach

Second, not only there is a peak using our new algorithm, but also that the peak can be approximately estimated. As we have discussed in Section 2, 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% and the corresponding accuracy number is 59.1%. The cutoff point found out is very close to the true peak point shown in the figure.

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 with the significant 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 at the current stage, we have not yet explored to use 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 the 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 hypothesized transcription, regenerated the lattices for all the candidate utterances, determined and selected the new hypothesized transcriptions, and retrained the AM with the new hypothesizes data. We achieved 59.32% accuracy, which is 0.2% better than the first iteration. If we use the whole (100%) candidate set with *true* transcriptions, we can obtain the upper bound which is 61.06%. The dotted red curve and the dashed green curve in Figure 1 compare the results using our proposed approach with one and two iterations respectively. It can be seen that the second iteration is slightly better than the first one.

4. SUMMARY AND CONCLUSIONS

We have described a new semi-supervised learning algorithm for improving acoustic models. The core idea of our algorithm is to determine and select the transcriptions based on both the utterance itself and acoustically similar other utterances. We approximate the effect an utterance and transcription pair will cause on the performance of the whole dataset with a globally defined expected entropy reduction by using confusion pairs observed between lattices. The effectiveness of our algorithm was demonstrated using the directory assistance data recorded under the real usage scenarios. The experiments indicated that our algorithm has better ability to identify the good hypothesis and utterances to be used for training the AM and to automatically identify the cutoff point.

There are many areas to improve along this line of research. For example, we have not utilized hypothesis other than the top one in our current algorithm and experiments, and the approximation we have made is rather crude. We tribute all these to the future work.

5. 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.

6. REFERENCES

[1] F. Wessel, K. Macherey, and R. Schluter, "Using word probabilities as confidence measures," in *Proc. of ICASSP*, 1998, pp. 225–228.

[2] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. of Eurospeech*, 1999, pp. 2725–2728.

[3] D. Charlet, "Confidence-measure-driven unsupervised incremental adaptation for hmm-based speech recognition," in *Proc. of ICASSP*, 2001, pp. 357–360.

[4] P. J. Moreno and S. Agarwal, "An experimental study of em based algorithms for semi-supervised learning in audio classification," in *Proc. of ICML-2003 Workshop on Continuum from Labeled to Unlabeled Data*, 2003.

[5] R. Zhang and A. I. Rudnicky, "A new data selection approach for semi-supervised acoustic modeling," in *Proc. of ICASSP*, 2006, pp. 421–424.

[6] 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.