ACTIVE AND UNSUPERVISED LEARNING FOR AUTOMATIC SPEECH RECOGNITION

Giuseppe Riccardi and Dilek Hakkani-Tür[*]

AT&T Labs-Research,
180 Park Avenue, Florham Park, NJ, USA
{dsp3, dtur}@research.att.com

ABSTRACT

State-of-the-art speech recognition systems are trained using human transcriptions of speech utterances. In this paper, we describe a method to combine active and unsupervised learning for automatic speech recognition (ASR). The goal is to minimize the human supervision for training acoustic and language models and to maximize the performance given the transcribed and untranscribed data. Active learning aims at reducing the number of training examples to be labeled by automatically processing the unlabeled examples, and then selecting the most *informative* ones with respect to a given cost function. For unsupervised learning, we utilize the remaining untranscribed data by using their ASR output and word confidence scores. Our experiments show that the amount of labeled data needed for a given word accuracy can be reduced by 75% by combining active and unsupervised learning.

1. INTRODUCTION

State-of-the-art speech recognition systems require transcribed utterances for training. Active learning aims at reducing the number of training examples to be labeled by inspecting the unlabeled examples, and selectively sampling the most *informative* ones with respect to a given cost function[1]. The goal of the active learning algorithm is to select the examples for labeling which will have the largest performance improvement. Unsupervised learning aims at utilizing unlabeled examples, to train, augment, or adapt systems. In this paper, we describe a new method for reducing the transcription effort for training in ASR, by combining active and unsupervised learning. Selective sampling involves automatically labeling each word of the utterance with a confidence score. We use the lattice output of a speech recognizer, which was initially trained on a small set of transcribed data, to compute word confidence scores. We then selectively sample the utterances to be transcribed, using utterance confidence scores computed from the word confidence scores. In addition to this, we exploit the utterances that were not selected for transcription using their ASR output and word confidence scores. Since no human-intervention is employed to use this data, it is called unsupervised learning [2, 3, 4, 5, 6, 7]. In Section 2, we describe our algorithms for active and unsupervised learning and how we combine them. In Section 3, we describe our experiments and results.

2. APPROACH

In this section, we present the active and unsupervised learning algorithms and their combination, in order to reduce the amount of transcription necessary to train language and acoustic models for ASR. The key concept for these algorithms is using word and utterance confidence scores. These scores are used for utterance selection in active learning, and probability estimation in unsupervised learning. In the following sections, we describe the main active and unsupervised learning algorithms, as well as their combination, and then how we compute word and utterance confidence scores from lattice output of ASR.

2.1. Active Learning

Inspired by the certainty-based active learning methods to reduce the transcription effort [8], we select the examples that we predict that the speech recognizer has misrecognized and give them to human labelers for transcription. We use these transcribed utterances for training acoustic and language models and leave out the ones that we predict the recognizer has recognized correctly or are not *informative*.

The first step of the algorithm is the training of initial language and acoustic models, using a small set of transcribed data, S_t. Using these models, we compute the speech utterances confidence scores and predict which candidate utterances are recognized (in)correctly [6]. The utterances are ranked according to their estimated correctness (the higher the score the higher the rank order) We then add the transcribed utterances to S_t and exclude them from S_u. This step is iterated as long as there are additional untranscribed utterances and the algorithm is halted if the word accuracy

[*] The authors are listed in alphabetical order.

on the development development set has converged. The algorithm for active learning is as follows:

1. Train acoustic and language models, AM_i and LM_i, for recognition, using S_t (i is the iteration number).
2. Recognize the utterances in set S_u using AM_i and LM_i, and compute the confidence scores for all the words.
3. Compute confidence scores of utterances.
4. Select k utterances which have the smallest confidence scores from S_u, and transcribe them. Call the new transcribed set as S_i.
5. $S_t = S_t \bigcup S_i$; $S_u = S_u - S_i$.
6. Stop if word accuracy has converged, otherwise go to Step 1.

In order to make better decisions in the future selections with respect to the labeling cost, k should be 1. However, for efficiency reasons in retraining, it is usually set higher.

2.2. Unsupervised Learning

Unsupervised learning aims to exploit untranscribed data to either bootstrap a language model or in general improve upon the model trained from the transcribed set of training examples. The core problem of unsupervised learning is the estimation of the error signal. In the case of language modeling the error signal is the noise on the event counts. Even in the simple case of n-gram language modeling the n-gram counts in the presence of noise are very unstable. In Figure 1 we show a scatter a plot of n-gram relative frequncies as estimated from clean (true) transcriptions of speech (x axis) and estimated from noisy transcriptions (y axis), in this case the ASR output. Figure 1 shows how the error variance increases for rare n-grams.

In standard n-gram estimation we count the occurrences of n-tuples $C(w_1^n)$, where w_1^n is the word n-tuple $w_1, w_2, \ldots, w_n$. In unsupervised learning the nature of the information is noisy and the n-gram counts are estimated from two synchronized information channels, the speech utterance hypothesis and the error signal. For each word w_i we estimate the probability of being correctly decoded c_i ($= 1 - e_i$, where e_i is the error probability), that is its confidence score. The bidimensional channel is then represented as a sequence of n-tuples of symbol pairs $(w_1^n, c_1^n) = (w_1, c_1)(w_2, c_2), \ldots, (w_n, c_n)$. The n-gram counts in presence of noise can be computed by marginalizing the joint channel counts:

$$C_{UL}(w_1^n) = \sum_{x \in \mathcal{T}} c_x \delta_{w_1^n}(x) \qquad (1)$$

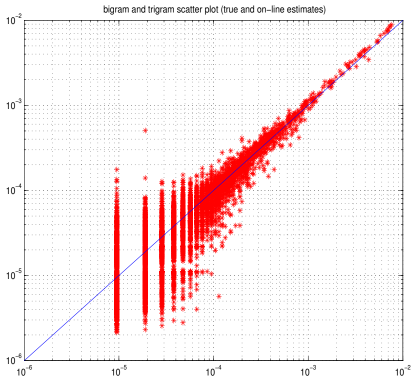

Fig. 1. Log-Log scatter plot of n-gram relative frequencies estimated from clean (x-axis) *vs* noisy (y-axis) speech utterance transcriptions.

where c_x is the confidence score for the n-tuple x in the noisy spoken utterance transcriptions $\mathcal{T}$ and $\delta_{w_1^n}(x)$ is the indicator function for the n-tuple w_1^n. The confidence score of the n-tuple w_1^n can be computed by geometric or arithmetic means or max and min over the n-tuple of word confidence scores c_1^n.In this paper we take the simplest approach and compute $c_{w_1^n} = c_n$. The equation 1 can be rewritten as a function of the error probability e_n:

$$C_{UL}(w_1^n) = C(w_1^n) - \sum_{x \in \mathcal{T}} e_x \delta_{w_1^n}(x) \qquad (2)$$

This equation shows the relation between the count estimates with and without error signal, $C_{UL}(w_1^n)$ and $C(w_1^n)$, respectively.

2.3. Combining Active and Unsupervised Learning

In order to combine active and unsupervised learning, we modify the first step of the active learning algorithm as following, where we have shown the additions in boldface fonts:

1. Train acoustic and language models, AM_i and LM_i, for recognition, using S_t (i is the iteration number), **and the ASR output of the utterances in S_u, as well as their word confidence scores**.

We train language models using the transcribed utterances, and the ASR output for the untranscribed utterances as explained in Section 2.2. The n-gram counts $C_{AL-UL}(w_1^n)$ from human transcribed (via Active Learning) and ASR transcribed speech utterances are computed in the following way:

$$C_{AL-UL}(w_1^n) = C_{AL}(w_1^n) + C_{UL}(w_1^n) \qquad (3)$$

The acoustic models can be trained by using all the data in a similar fashion [3].

3. EXPERIMENTS AND RESULTS

We performed a series of experiments to verify that word confidence scores can be used to identify correctly recognized words, utterance confidence scores can be used to select more informative utterances to transcribe, and automatic speech recognition accuracy can be improved by exploiting untranscribed data. For all these experiments, we used utterances from the *How May I Help You?*$^{\mathrm{SM}}$ speech database [13]. The language models used in all our experiments are trigram models based on Variable Ngram Stochastic Automata [14].

3.1. Training and Test Data

In the *How May I Help You?*$^{\mathrm{SM}}$ speech database there are two distinct data collections. The first is from human-human conversations (8K utterances and 300K word tokens) and consists of responses to the initial prompt. The second is from human-machine dialogs (28K and 318K word tokens) and consists of users' responses to all system prompts (e.g. greeting and confirmation prompts). The test data consists of 1,000 utterances (10K words) from the human-machine data collection. In all the experiments presented in this paper, we kept the triphone context acoustic model fixed The acoustic model has been trained using utterances from human-human conversations, and off-the-shelf telephone speech corpora. The training data for the acoustic model does not overlap with our additional training data.

3.2. Confidence Score Computation

We extract word confidence scores from the lattice output of ASR. The algorithm is based on the *pivot* alignment for strings in the word lattice. A detailed explanation of this algorithm and the comparison of its performance with other approaches is presented in [11]. To check how good the word confidence scores are in distinguishing the correctly recognized and misrecognized words, we considered a binary classification problem, where we used the confidence scores of the words, as well as a threshold for the decision. We classified each word as correctly recognized if that word has a confidence score higher than the threshold, and as misrecognized otherwise. The false rejection and false acceptance rates achieve the equal error rate at 22% on the test set.

3.3. Active and Unsupervised Learning

For active learning in ASR, we trained an initial language model, using the initial set utterances (8K) from human-human interactions. Using this model, we then generated lattices and pivot alignments for our additional training data, and computed the confidence scores for words and utterances, as described in Section **??**. Using the confidence scores for utterances, we sorted the utterances in an order according to *informativeness* for ASR (generating the selectively sampled order). We incrementally trained language models only, by using the top k utterances from the randomly sampled and selectively sampled orders, and generated learning curves for vocabulary size and word accuracy, which are presented in Figures 2 and 3. In this experiment, we used the arithmetic mean of the word confidence scores as utterance confidence scores, which gave the best results in our case. The active learning algorithm is independent of the way we compute the confidence scores. In our experiments, we also used the normalized utterance likelihood as a sampling criterion, and it gave inferior performance.

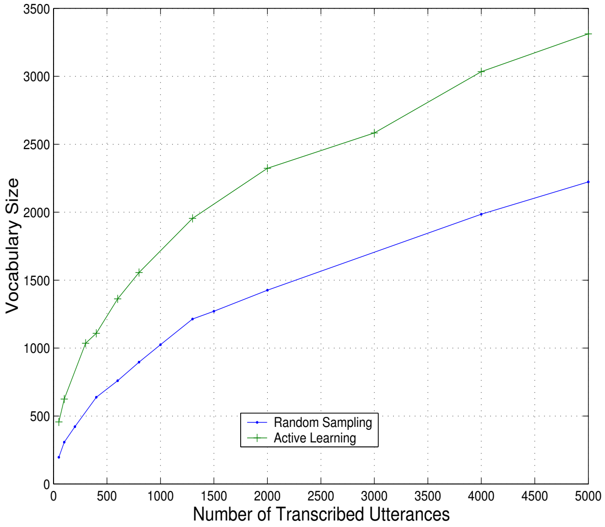

Fig. 2. Vocabulary size growth with random and selective sampling.

Figure 3 also shows a learning curve when we combined active and unsupervised learning. For this curve, we took the selectively sampled order, and we added the transcriptions of the top k utterances, and the automatic speech recognizer output of the rest of the utterances to our training set for the language models. The recognizer output was generated using the initial language model, and contained word confidence scores. Figure 3 shows the word accuracy learning curves when we exploited additional untranscribed data by combining active learning with unsupervised learning.

From these curves, we see that active learning is effective in reducing the need for labeled data (for a given word accuracy). For example, to achieve 66.5% word accuracy with random sampling, we needed to transcriptions for 4,000 utterances, however, we are able to achieve this accuracy transcribing only around 2,500 utterances. This shows that we can achieve the same performance by transcribing

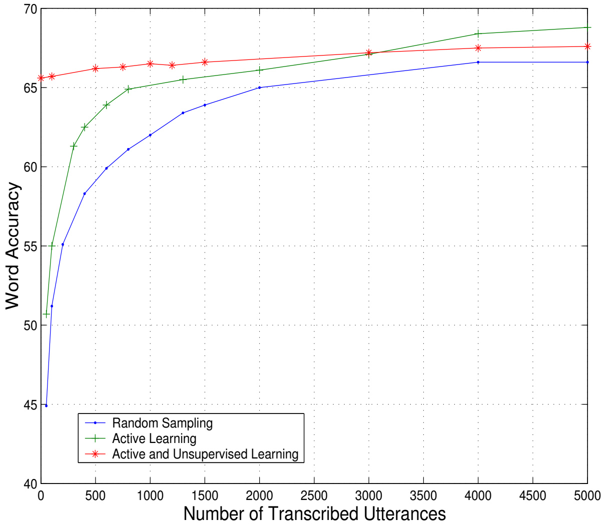

Fig. 3. Word Accuracy Learning Curves.

38% fewer utterances, when we use active learning. In addition to this, we get a huge improvement using the untranscribed data, at the very initial points. For example, when we use the combination of active and unsupervised learning, we get 66.5% by using only 1,000 transcribed utterances instead of 4,000 transcribed utterances, that is 75% less utterances than random sampling. As we have more transcribed data, the improvement using untranscribed data gets less and active learning takes over. The combination always results in higher word accuracy than random sampling, by 2-3% points.

4. CONCLUSIONS

We described new methods for reducing the amount of labeled training examples by selectively sampling the most informative subset of data for transcription using lattice based confidence measures, and exploiting the rest of the data, that has not been transcribed, by using their ASR output and word confidence scores. We have empirically shown that it is possible to detect utterances which have little new information when added to an initial set of utterances. In addition to this, we have shown that it is possible to exploit the untranscribed data, and we can achieve the same word accuracy results using 75% less data by combining active and unsupervised learning.

5. REFERENCES

[1] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, pp. 201–221, 1994.

[2] Thomas Kemp and Alex Waibel, "Unsupervised training of a speech recognizer using tv broadcasts," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP-98)*, 1998, pp. 2207–2210.

[3] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *Proceedings of the Broadcast News Transcription and Understanding Workshop*. 1998, pp. 301–305, Morgan Kaufman.

[4] Thomas Kemp and Alex Waibel, "Learning to recognize speech by watching television," *IEEE Intelligent Systems*, pp. 51–58, 1999.

[5] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[6] Roberto Gretter and Giuseppe Riccardi, "On-line learning of language models with word error probability distributions," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001.

[7] Andreas Stolcke, "Error modeling and unsupervised language modeling," in *Proceedings of the 2001 NIST Large Vocabulary Conversational Speech Recognition Workshop*, Linthicum, Maryland, 2001.

[8] D.D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. of the* 11^{th} *International Conference on Machine Learning*, 1994, pp. 148–156.

[9] R.C. Rose, B.H.Juang, and C.H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1995, pp. 281–284.

[10] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[11] Dilek Hakkani-Tür and Giuseppe Riccardi, "A general algorithm for word graph matrix decomposition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003.

[12] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. of* 7^{th} *European Conference on Speech Communication and Technology*, 2001, pp. 2105–2108.

[13] A. Gorin, J.H. Wright, G. Riccardi, A. Abella, and T. Alonso, "Semantic information processing of spoken language," in *Proc. of ATR Workshop on Multi-Lingual Speech Communication*, 2000.

[14] G. Riccardi, R. Pieraccini, and E. Bocchieri, "Stochastic automata for language modeling," *Computer Speech and Language*, vol. 10, pp. 265–293, 1996.