Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models

Thomas Drugman, Janne Pylkkönen, Reinhard Kneser

Amazon

drugman@amazon.com, jannepyl@amazon.com, rkneser@amazon.com

Abstract

The goal of this paper is to simulate the benefits of jointly applying active learning (AL) and semi-supervised training (SST) in a new speech recognition application. Our data selection approach relies on confidence filtering, and its impact on both the acoustic and language models (AM and LM) is studied. While AL is known to be beneficial to AM training, we show that it also carries out substantial improvements to the LM when combined with SST. Sophisticated confidence models, on the other hand, did not prove to yield any data selection gain. Our results indicate that, while SST is crucial at the beginning of the labeling process, its gains degrade rapidly as AL is set in place. The final simulation reports that AL allows a transcription cost reduction of about 70% over random selection. Alternatively, for a fixed transcription budget, the proposed approach improves the word error rate by about 12.5% relative.

Index Terms: speech recognition, active learning, semi-supervised training, data selection

1. Introduction

This paper aims at the problem of identifying the best approach for jointly selecting the data to be labelled, and maximally leveraging the data left as unsupervised. Our application targets voice search as used in various Amazon products. Because speech data transcription is a time-consuming and hence costly process, it is crucial to find an optimal strategy to select the data to be transcribed via active learning. In addition, the unselected data might also be helpful in improving the performance of the ASR system by semi-supervised training. As will be shown in this paper, such an approach allows to reduce the transcription cost dramatically while enhancing the customer's experience.

Active Learning (AL) refers to the task of minimizing the number of training samples to be labeled by a human so as to achieve a given system performance [1]. Unlabeled data is processed and the most informative examples with respect to a given cost function are then selected for human labeling. AL has been addressed for ASR purpose across various studies, which mainly differ by the measure of informativeness used for data selection. First attempts were based on so-called confidence scores [2] and on a global entropy reduction maximization criterion [3]. In [4], a committee-based approach was described. In [5], a min-max framework for selecting utterances considering both informativeness and representativeness criteria was proposed. This method was used in [6] together with an N-best entropy based data selection. Finally, the study in [7] found that HMM-state entropy and letter density are good indicators of the utterance informativeness. Encouraging results were reported from the early attempts [2, 3] with a 60% reduction of the transcription cost over Random Selection (RS).

In this paper, we focus on conventional confidence-based AL as suggested in [2], although other studies [3, 6, 7] have shown some improvement over it. It is however worth highlighting that the details of the baseline confidence-based approach were not always clearly described, and that subsequent results were not in line with those reported in [2]. First, various confidence measures can be used in ASR. A survey of possible confidence measures is given in [8] and several techniques for confidence score calibration have been developed in [9]. Secondly, there are various possible ways of selecting data based on the confidence scores.

Semi-Supervised Training (SST) has also recently received a particular attention in the ASR literature. A method combining multi-system combination with confidence score calibration was proposed in [10]. A large-scale approach based on confidence filtering together with transcript length and transcript flattening heuristics was used in [11]. A cascaded classification scheme based on a set of binary classifiers was proposed in [12]. A reformulation of the Maximum Mutual Information (MMI) criterion used for sequence-discriminative training of Deep Neural Networks (DNN) was described in [13]. A shared hidden layer multi-softmax DNN structure specifically designed for SST purpose was proposed in [14]. The way unsupervised data[1] is selected in this paper is inspired from [11], as it is based on confidence filtering with possible additional constraints on the length and frequency of the transcripts.

This paper aims at addressing the following questions whose answer is left either open or unclear with regard to the literature: *i)* Do more sophisticated confidence models help improving data selection? *ii)* Is AL also beneficial for LM training, and if so to which extent? *iii)* How do the gains of AL and SST scale up when more and more supervised data is transcribed? *iv)* Are the improvements similar after cross-entropy and sequence-discriminative training of the DNN AM?

In most of existing AL and SST studies (e.g. [2, 3, 6, 7, 13, 14]), the Word Error Rate (WER) typically ranges between 25 and 75%. The baseline model in the present work has a WER of about 12.5%, which makes the application of AL and SST on an industrial task even more challenging.

The paper is structured as follows. Section 2 presents the approach studied throughout our experiments. Experimental results are described in Section 3. Finally, Section 4 concludes the paper.

---

[1] In this paper, "unsupervised data" refers to transcriptions automatically produced by the baseline ASR. In literature, training with automatic transcriptions produced by a supervised ASR system is sometimes referred to as "semi-supervised training", but we reserve the latter term to situations where both manual and automatic transcriptions are used together.

2. Method

Our method relies heavily on confidence-based data selection. Because confidence scores play an essential role, several confidence models have been investigated. They are described in Section 2.1. The technique of data selection is presented in Section 2.2. Details about AM and LM training are then provided respectively in Sections 2.3 and 2.4.

2.1. Confidence modeling

As mentioned in the introduction, there are various confidence measures available [8, 9, 15]. First of all, confidence measures can be estimated at the token and utterance levels. The conventional confidence score at the token level is the token posterior from the confusion network [8]. It is however in practice a poor estimate of the actual probability of the token being correct, and is therefore lacking interpretability. This was addressed in [9] by calibrating the scores using a maximum entropy model, an artificial neural network, or a deep belief network.

In this paper, confidence score normalization is performed to match the confidences with the observed probabilities of words being correct, using one of the two following methods: a piecewise polynomial which maps the token posteriors to confidences, or a linear regression model with various features such as the token posteriors, the word accuracy priors, the number of choices in the confusion network, and the number of token nodes and arcs explored. These two models are trained on an in-domain held-out data set.

As our data selection method processes utterances, it is necessary to combine the scores from the various tokens to get a single confidence measure at the utterance level. Conventional approaches encompass using an arithmetic or geometrical mean rule. In addition, we have also considered training a Multi-Layer Perceptron (MLP) to predict either the WER or the Sentence Error Rate (SER). The MLP took as input a vector consisting of statistics of the tokens: number of tokens, min/max and mean values of their posteriors.

2.2. Data Selection

Because the DNN we use for AM is a discriminative model, the selection of supervised data for AM purpose consists in maximizing the informativeness of the chosen utterances. Intuitively, this translates to selecting utterances with low confidence scores. Different settings of the confidence filter will be investigated in our experiments. Besides, we consider also filtering out too short utterances.

The selection of unsupervised data requires to find a balance between the informativeness and the quality of the automatic transcripts. This latter aspect imposes to retain only high confidence scores, as errors in the transcripts can be harmful to the training (particularly if it is sequence-discriminative [13]). As suggested in [11], utterance length and frequency filtering are additionally applied to flatten the data.

2.3. AM training

Our AM is a conventional DNN [16] made of 4 hidden layers containing 1536 units each. A context-dependent GMM is first trained using the Baum-Welch algorithm and PLP features. The size of the triphone clustering tree is about 3k leaves. The GMM is used to produce the initial alignment of the training data and define the DNN output senones. Our target language in this study is German, but we decided to apply transfer learning [17] by initializing the hidden layer weights from a previously-trained English DNN. The output layer was initialized with random weights. The input features are 32 standard Mel-log filter bank energies, spliced by considering a context of 8 frames on each side, therefore resulting in 544 dimensional input features.

The training consists of 18 epochs of frame-level cross-entropy (XE) training followed by boosted Maximum Mutual Information (bMMI) sequence-discriminative training [18]. The Newbob algorithm is used as Learning Rate (LR) scheduler during XE training. The learning rate for bMMI was optimized using a held-out development set. The resulting DNN is used to re-align the data and the same procedure of DNN training starting from transfer learning is applied again. The baseline model on the 50 initial hours was obtained in this way. For the next models which ingest additional supervised and/or unsupervised data, the baseline model is used to get the alignments, and the training procedure starting from transfer learning is performed.

2.4. LM training

Our LM is a linearly interpolated trigram model consisting of 9 components. The most important one (with interpolation weight > 0.6) is trained on the selected supervised and unsupervised data. For the remaining components, we consider a variety of Amazon catalogue and text search data relevant for the voice search task. All component models are 3-gram models trained with modified Kneser-Ney smoothing [19]. The interpolation parameters are optimized on a held-out development set. The size of the LM is finally reduced using entropy pruning [20].

3. Experiments

The aim of our experiments is to simulate the possible gains obtained by AL and SST for a new application. For this simulation, we had about 600 hours of transcribed voice search data in German at our disposal. From this pool, 50 hours are first randomly selected to build the baseline AM and LM. These models are then used to decode the remaining 550 hours. The confidence models described in Section 2.1 and previously trained on a held-out set are employed so that each utterance in the 550h selection pool is assigned one confidence score (per confidence model). From the selection pool, the supervised data is selected first via conventional RS or via AL. Utterances which were left over are considered as unsupervised data for SST. The evaluation is carried out on a held-out dataset of about 8 hours of in-domain data. A speaker overlap with the training set is possible but the large number of speakers diminishes its potential effect. Our target metric is the standard WER.

In the next sections the results of the experiments are presented. Section 3.1 investigates the influence of the confidence model on data selection. The impact of AL and SST on both the AM and the LM is studied in Sections 3.2 and 3.3. Lastly, Section 3.4 simulates the final gains on a new ASR application.

3.1. Confidence modeling

Various confidence models including a normalization of the token posteriors and an utterance-level calibration, as described in Section 2.1, have been tried for data selection. For each confidence model, the confidence filter settings have been optimized as will be explained in Section 3.2. Unfortunately, our results did not indicate any AL improvement by using more sophisticated confidence models. Only marginal (below 2% relative WER) differences not necessarily consistent across the experiments were observed. Our explanation is two-fold: First, the ranking across the utterances in the selection pool is not sub-

stantially affected by the different models. Second, even when the ranking is altered, the informativeness of the switched utterances is probably comparable, therefore not leading to any dramatic difference in recognition performance.

The rest of this paper therefore employs a simple confidence model: a polynomial is used to map the token posteriors to the observed word probabilities, which are then combined by geometrical mean. The distribution of these scores over the 550h selection pool is shown in Figure 1. Note that the various peaks in the high confidences are due to a dependency on the hypothesis length. As can be seen, the baseline model is already rather good: respectively 11.6, 19.0 and 24.0% of the utterances have a confidence score lower than 0.5, 0.7 and 0.8.
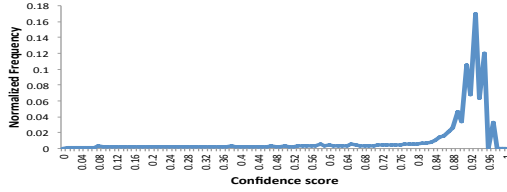

Figure 1: Histogram of the standard confidence scores.

3.2. Impact on the AM

In this section, we focus on the impact of AL and SST purely on the AM. The LM and the vocabulary are therefore fixed to that of the baseline. For both supervised and unsupervised data selection, our approach relies on applying a filter to the confidence scores where data is selected if the confidence score is between some given lower and upper bounds.

3.2.1. Active learning only

In a first stage, we optimized the filter used for supervised data selection. We varied the lower filter bound in the [0-0.1] range in order to remove possibly uninformative out-of-domain utterances. The upper bound was varied in the [0.4-0.9] range, leading to a total of 20 filters. The resulting AMs were analyzed on the development set. The main findings were that as long as the lower bound does not exceed 0.05 and the upper bound does not exceed 0.8 (which corresponds to the beginning of the main mode in Figure 1), the results were rather similar (with differences lower than 1% relative). It seems to be important, though, not to go beyond 0.8 as this would strongly compromise the informativeness of the selected utterances. In addition, we have tried to apply utterance length filtering in cascade with the confidence-based selection. This operation however did not turn out to provide any gain.

Based on these observations, we have used the [0-0.7] confidence filter for AL data selection. When 100h of supervised data was added to the baseline, this technique reduced the WER by about 2% relative over the RS scheme.

3.2.2. Including unsupervised data

In a second stage, we optimized the method for selecting the unsupervised data. On top of the 50h baseline set and the 50h of AL data (selected as mentioned in Section 3.2.1) we added unsupervised data selected according to different confidence filters and analyzed again the AM performance after XE training on the development set.

Our attempts to integrate utterance length and frequency filtering as in [11] were not conclusive as no significant gains were obtained. We also remarked a slight degradation if the upper
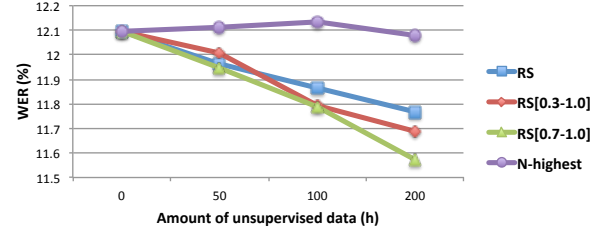

Figure 2: Benefits of unsupervised data on a XE-trained AM.

bound for confidence filtering does not reach the limit of 1.0. We therefore focused on pure confidence filtering with an upper bound of 1.0 in the remainder of our experiments. The plot in Figure 2 compares 4 techniques of unsupervised data selection: unfiltered random sampling (*RS*), confidence filtering using two different confidence filters (*RS[0.3-1.0]* and *RS[0.7-1.0]*), and choosing the sentences with the highest confidence scores (*N-highest*). We obtained the best results with the [0.7-1.0] confidence filter. The poor performance of the N-highest scores approach can be explained by the fact that it just adds high confidence utterances which contain little new information. On the other hand, with a low lower bound of the confidence filter (as in [0.3-1.0] or *RS*) the label quality becomes worse and the results also degrade. A remarkable fact is that the more unsupervised data, the better the performance of the AM. The addition of 200h of unsupervised data yielded an improvement of 4.5% relative. The same experiment was replicated with 100h of AL data, and the conclusions remained similar, except that the gain reached 3.5% (and not 4.5%) this time.

3.3. Impact on the LM

The most important component of the interpolated LM is the one trained on transcriptions of the in-domain utterances. In this section we study the impact of different methods to select in-domain data and add it to this component on top of the 50h of the baseline model. All other LM components are kept constant. We consider three data pools from which training data could be taken: supervised data from the 100h AL data pool which was selected using the [0-0.7] confidence filter as described in Section 3.2.1, supervised data from the complete pool of 550h, and unsupervised data from the same 550h pool, taken from the first hypothesis of the ASR results of the baseline model.

3.3.1. Perplexity results

In a first experiment we calculated perplexities when an increasing amount of data was added to the LM. Since perplexity values are hard to compare when using different vocabularies, we kept the vocabulary fixed to that of the baseline. The dotted lines in Figure 3 show the perplexities if data is randomly sam-
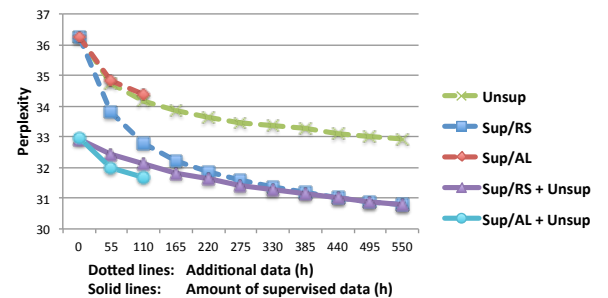

Figure 3: LM perplexity for different types of data.

pled from the supervised data (*Sup/RS*), if the data is sampled from the recognition results (*Unsup*), and if the data is sampled from the the AL data pool (*Sup/AL*). It can be seen that adding more application data improves the model irrespective of the source. Already just adding unsupervised data gives a big perplexity reduction from 36.3 to 33.0. However, there is a significant gap between the supervised (*Sup/RS*) and the unsupervised (*Unsup*) case. Adding just the AL data does not perform as well as random sampling from the complete pool. On one hand, in this case the label quality is higher compared to the unsupervised data but on the other hand, due to the selection process, the data is no longer representative to the application. Contrary to the AM, which is discriminatively trained, the LM is a generative model which in general is much more vulnerable for missing representativeness.

In the next experiments, shown as solid lines in Figure 3, we combine supervised and unsupervised data with the goal to overcome the bias in the data and to make the best use of all the data. Supervised data was again selected either by RS (*Sup/RS + Unsup*) or by AL (*Sup/AL + Unsup*) but in addition, all the remaining data of the 550h data pool were used in training as unsupervised data. This way we always use the complete data and thus maintain the representativeness. The beginning of the curves correspond to 550h unsupervised data. In the case of *Sup/RS + Unsup* it drops constantly to final value of 550h supervised data. Contrary to the previous experiment, when applying AL to select the training data (*Sup/AL + Unsup*), we no longer suffer from a bias of the data and the model performs even slightly better than RS.

3.3.2. Recognition results

It is well known that gains in perplexity do not always correspond to WER improvements. We therefore ran recognition experiments using the LMs from Section 3.3.1. Since it is beneficial to the models we always added the unsupervised data on top of the supervised data. The AM was kept fixed to the baseline. As we were no longer restricted by the perplexity measure, we also updated the vocabulary according to the selected supervised training data in these experiments. The results in Figure 4 show that the improvements in perplexity are also reflected in a better WER even though part of the improvements might also be due to the increased vocabulary coverage. It is interesting to observe that, when adding 100 hours of supervised data, the gains for AL are much higher than for RS. In total, the impact of AL combined with SST on LM is outstanding: after 100h of transcribed data, the gain over the RS baseline reaches 5.3% relative. It is also worth emphasizing that 100h AL and roughly 400h RS are equivalent in terms of LM performance.
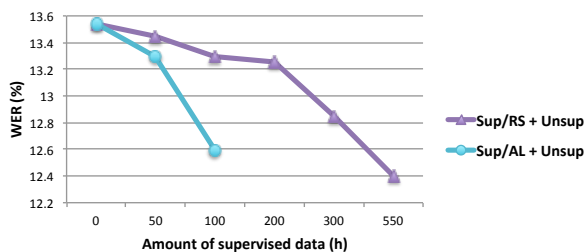

Figure 4: ASR results with updated LM and vocabulary.

3.4. Final results

Finally, we simulate the improvements that would be yielded in a new application by applying confidence-based AL and SST

to both the AM and LM. We considered the different LMs as suggested in Section 3.3. For AM building, we limited the unsupervised set to 200h across our experiments. For XE training, SST was applied, following the findings from Section 3.2.2. For sequence-discriminative bMMI training, it is known that possible errors in the transcripts can have a dramatic negative influence on the quality of the resulting AM [13]. Therefore, two strategies were investigated: *i)* considering the aggregated set of supervised and unsupervised data for bMMI training; *ii)* discard any unsupervised data and only train on the supervised set. Our results indicate that the inclusion of unsupervised data led to a degradation of about 2.5%, and this despite the relatively high lower bound used in the confidence filter (0.7). The first strategy was therefore used in the following.
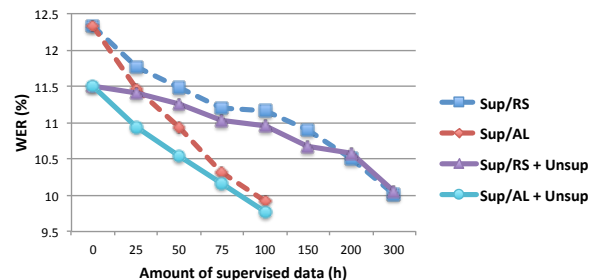

Figure 5: Final simulation: both the AM and LM are updated.

Figure 5 shows the final simulation results after bMMI training. It is worth noting that the results obtained after XE training were very much in line and led to very similar improvements. Two main conclusions can be drawn from this graph. First, the unsupervised data is particularly important at the very beginning, where it allows a 6.8% relative improvement. Nevertheless, the gains of SST vanish as more supervised data is collected. In the AL case, the advantage from SST almost completely disappears after 100h of additional supervised data. Secondly, AL carries out significant improvements over RS. It can be seen that the WER obtained with 100h of AL is comparable (even slightly better) to that using 300h of RS data, hence resulting in a reduction of the transcription budget of about 70%. Alternatively, one can observe that, for a fixed transcription cost of 100h, AL achieves an appreciable WER reduction of about 12.5% relative over the range of added supervised data.

4. Conclusions

This paper aimed at simulating the benefits of AL and SST in a new ASR application by applying confidence-based data selection. More sophisticated confidence models have been developed, but they did not provide any gain for training data selection for AL. Regarding AM training, AL alone was found to yield a 2% relative improvement. Combining it with SST turned out to be essential, especially when the amount of supervised data is limited. Adding 200h of unsupervised data to 50h of AL gave a 4.5% gain on the AM trained by cross-entropy. On the contrary, any unsupersived data was harmful to sequence-discriminative bMMI training. Beyond these improvements on the AM, combining AL and SST allowed a significant improvement (about 5%) of the LM. Our final results indicate that applying AL to both AM and LM provides an encouraging 70% reduction of the transcription budget over RS, and these gains seem to scale up rather well as more and more utterances are transcribed.

5. References

[1] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

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

[3] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language*, vol. 24, pp. 433–444, 2010.

[4] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," *ICASSP*, pp. 4350–4353, 2010.

[5] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and repsentative examples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1936–1949, 2014.

[6] N. Itoh, T. Sainath, D. Jiang, J. Zhou, and B. Ramabhadran, "N-best entropy based data selection for acoustic modeling," *ICASSP*, pp. 4133–4136, 2012.

[7] T. Fraga-Silva, J. Gauvain, L. Lamel, A. Laurent, V. Le, and A. Messaoudi, "Active learning based data selection for limited resource stt and kws," *Interspeech*, pp. 3159–3162, 2015.

[8] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[9] D. Yu, J. Li, and L. Deng, "Calibration of confidence measures in speech recognition," *IEEE Transactions on Audio, Speech and Language*, vol. 19, no. 8, pp. 2461–2473, 2011.

[10] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised gmm and dnn acoustic model training with multi-system combination and confidence re-calibration," *Interspeech*, p. 23602364, 2013.

[11] O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," *Interspeech*, p. 20832087, 2014.

[12] S. Li, Y. Akita, and T. Kawahara, "Discriminative data selection for lightly supervised training of acoustic model using closed caption texts," *Interspeech*, p. 35263530, 2015.

[13] V. Manohar, D. Povey, and S. Khudanpur, "Semi-supervised maximum mutual information training of deep neural network acoustic models," *Interspeech*, p. 26302634, 2015.

[14] H. Su and H. Xu, "Multi-softmax deep neural network for semi-supervised training," *Interspeech*, p. 32393243, 2015.

[15] Z. Bergen and W. Ward, "A senone based confidence measure for speech recognition," *Eurospeech*, pp. 819–822, 1997.

[16] G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[17] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," *ICASSP*, pp. 8619–8623, 2013.

[18] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *Interspeech*, pp. 2345–2349, 2013.

[19] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," *ICASSP*, vol. 1, pp. 181–184, 1995.

[20] A. Stolcke, "Entropy-based pruning of backoff language models," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 270–274, 1998.