Acoustic Model Bootstrapping Using Semi-Supervised Learning

Langzhou Chen, Volker Leutnant

Amazon

langzhou@amazon.com,leutnant@amazon.de

Abstract

This work aims at bootstrapping the acoustic model training with small amount of the human annotated speech data and large amount of the unlabelled speech data for automatic speech recognition.The technologies of the semi-supervised learning were investigated to select the automatically transcribed training samples.Two semi-supervised learning methods were proposed: one is the local-global uncertainty based method which introduces both the local uncertainty from the current utterance, and the global uncertainty from the whole data pool into the data selection; the other is the margin based data selection, which selects the utterances near to the decision boundary through the language model tuning. The experimental results based on a Japanese far-field automatic speech recognition system indicated that the acoustic model trained by the automatically transcribed speech data achieved about 17% relative gain when the in-domain human annotated data was not available for initialization. While 3.7% relative gain was obtained when the initial acoustic model was trained by small amount of the in-domain data.

Index Terms: speech recognition, semi-supervised training

1. Introduction

Bootstrapping the acoustic model (AM) training for the automatic speech recognition (ASR) system building with small amount of the human annotated data is a challenging task, since the performance of the ASR system strongly relies on the size and the quality of the training speech. Semi-supervised learning (SSL) methods [1] [2] [3] [4] aim at training the ASR system with automatically transcribed data. It becomes an important research area since nowadays large amount of speech data can be collected with low cost, but the human annotation of the data is still expensive and time-consuming.

To select the automatically transcribed data for SSL, the confidence score was widely used to identify the reliable transcripts from the ASR output [5] [6]. Another type of data selection was not only based on the ASR output, but also the less accurate transcripts of the speech data, e.g. the close captions of the broadcast news data. Sometimes this kind of methods were referred to as lightly supervised training [7]. More complex data selection methods were also proposed in SSL data selection. In [8], multiple ASR systems were trained to automatically transcribe the speech data, and a cascade of the conditional random field models were used to combine the ASR hypotheses from different systems and judge the reliability of the automatically transcribed data. [9] proposed the global entropy reduction maximization (GERM) method. The utterances which caused the biggest global entropy reduction of the whole training data were selected. It achieved the balance between the informativeness of the selected samples and the size of the selected training data. As the deep learning dominating the research in speech domain, the new ways to make use of the unlabelled training data were proposed. The teacher-student models [10] [11] were used to train the ASR model to minimize the Kullback–Leibler distance between the output of the teacher model and the student model. This can be applied in either frame-wise level [10], or sequence level [12].

In this work, two different SSL data selection methods were proposed. The first method is the local-global uncertainty based method. The method simplified the SSL algorithm in [9]. The GERM method proposed in [9] is powerful to select the reliable and informative training samples from the automatically transcribed data pool. However, the calculation cost of GERM is high when the size of the data pool is big. Instead of calculating the global entropy reduction in the utterance level, this work broke down the utterance level uncertainty to word level using confusion network. Meanwhile, this work kept the idea of GERM to select the training samples by considering the global information. That says the uncertainty of each word in the ASR hypotheses was influenced by the similar samples in the data pool. This way, the calculation of the data selection was simplified and the advantages in original GERM method was still kept. The second method proposed in this work is the margin based SSL, which selects the samples close to the decision boundary. In the tasks of ASR, it is intractable to calculate the distance between the training sample and the decision boundary. This work proposed an alternative way to implement the margin based method. The language model (LM) tuning was used to adjust the decision boundary and select the training samples.

2. Semi-supervised Learning

The SSL methods take the ASR hypotheses as the transcripts of the speech utterance, and then re-train the ASR model using this automatically transcribed speech data. Since the automatically transcribed speech data was used in the ASR model training, the ASR errors may be introduced into the model training. On the other hand, if only the very reliable ASR hypotheses were used for training, the ASR system can not learn the new knowledge from the training samples. Thus the art of the SSL is always the balance between the reliability and the informativeness of the training samples.

2.1. SSL based on confidence score

Similar to the active learning (AL) based data selection, the confidence score based methods were also used in SSL data selection. However, in contrast to the AL, when the confidence methods were used in SSL, the training samples with high confidence were selected [6]. The problem of the confidence based SSL is that it is inclined to select the utterances which have already been recognized well. Thus the new knowledge learned by the ASR system is limited.

2.2. Local-global uncertainty based SSL

To address the problem of the confidence based methods, [9] proposed GERM based SSL. The basic idea is using not only the information of the current utterance, but also the information from the whole data pool to handle the data selection. The global entropy reduction was used to guide the data selection. Meanwhile to reduce the redundancy of the selected data, [9] assumed that every time an utterance was selected, the entropies of all the other utterances with the similar confusion patterns should be adjusted. This increased the complexity of the data selection process. In the experiments of [9], the data pool only contained 30k utterances. While in the state-of-the-art ASR systems, the amount of the training data is much bigger. For example, in this work, the size of the unlabelled data is several hundred times bigger than that in the experiments of [9]. This makes the calculation cost of the original GERM algorithm dramatically increased.

In this work, a local-global uncertainty based SSL data selection method was proposed. It attempted to keep the advantages of the method in [9], meanwhile simplified the data selection algorithm. In the proposed method, the SSL data selection was performed according to word level uncertainties which were calculated through word confusion networks [13]. The confusion network represents the ASR output in a sausage structure, which contains a number of alignment positions. In each alignment position, a set of mutually exclusive word hypotheses were defined, together with their posterior probabilities. Based on the confusion network, the proposed method defines the utterance uncertainty as the average of the word uncertainties in every alignment position in the sausage structure, i.e.

$$\mathbf{UNC}(u) = \frac{\sum_{t \in T_u} \mathbf{unc}(t,u)}{T_u} \tag{1}$$

where $\mathbf{UNC}(\mathbf{u})$ is the utterance uncertainty, $\mathbf{unc}(t,u)$ is the word level uncertainty in the alignment position t of utterance u, and T_u is the number of alignment positions in u.

Without considering the impacts from other utterances, the word level uncertainty can be calculated according to the entropy of the word hypotheses in the alignment position t, i.e.

$$\mathbf{unc_{local}}(t,u) = \sum_{w \in W_{t,u}} \mathbf{p}(w) \log(\mathbf{p}(w)) \tag{2}$$

where $W_{t,u}$ is the set of the word hypotheses in alignment position t, and $\mathbf{p}(w)$ was obtained from the posteriors of w in the confusion network. The uncertainty calculated by equation 2 only considers the local information for the current utterance. Thus it was defined as local uncertainty in this work.

The local uncertainty does not consider the information from the whole data pool. Simply using local uncertainty for data selection will end up with the results similar to the confidence based methods. To take into account the information from the whole data pool, the global uncertainty was introduced into the data selection, as shown in figure 1. Supposed that in the confusion network of utterance u, alignment position t, there is a word-pair a, b, where a is the best hypothesis and b is the second best with the most confusion. The idea of global uncertainty is looking for the data samples which contain the word-pair a, b in the data pool. If a sample in data pool was found with a as the best hypothesis and b as the second, it is a positive sample and reduces the uncertainty of the ASR hypotheses, since it is consistent with the current pattern. On the other hand, if a sample in data pool was found with b as the best hypothesis and a as the second, it is inconsistent with the current pattern. Thus it is a negative sample and increases the uncertainty of the hypotheses. The global uncertainty was calculated according to all the positive and negative samples which were observed in the data pool.

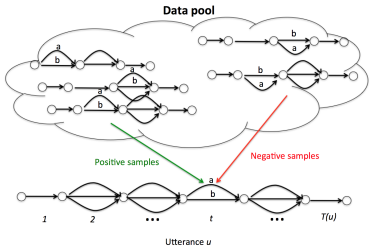

Figure 1: *Global uncertainty, from positive and negative samples*

Considering both local uncertainty and the global uncertainty, the word level uncertainty can be calculated as:

$$\mathbf{unc}(t,\hat{u}) = \mathbf{unc_{local}}(t,\hat{u}) * \max(0, 1 - \delta_{\mathbf{pos}}(t,\hat{u}) + \delta_{\mathbf{neg}}(t,\hat{u})) \tag{3}$$

where $\delta_{\mathbf{pos}}(t,\hat{u})$ and $\delta_{\mathbf{neg}}(t,\hat{u})$ are the weights from the positive samples and negative samples respectively, which can be defined as:

$$\begin{aligned}
\delta_{\mathbf{pos}}(t,\hat{u}) = & \sum_{u} \sum_{\tau \in T_u} \alpha \\
& * \exp(-\beta * \mathbf{kld}(\mathbf{p}(w_{1,t,\hat{u}}), \mathbf{p}(w_{2,t,\hat{u}}) : \mathbf{p}(w_{1,\tau,u}), \mathbf{p}(w_{2,\tau,u}))) \\
& * \mathbf{I}(w_{1,t,\hat{u}} = w_{1,\tau,u}, w_{2,t,\hat{u}} = w_{2,\tau,u}) \\
\delta_{\mathbf{neg}}(t,\hat{u}) = & \sum_{u} \sum_{\tau \in T_u} \alpha \\
& * \exp(-\beta * \mathbf{kld}(\mathbf{p}(w_{1,t,\hat{u}}), \mathbf{p}(w_{2,t,\hat{u}}) : \mathbf{p}(w_{2,\tau,u}), \mathbf{p}(w_{1,\tau,u}))) \\
& * \mathbf{I}(w_{1,t,\hat{u}} = w_{2,\tau,u}, w_{2,t,\hat{u}} = w_{1,\tau,u})
\end{aligned}$$

where $w_{1,t,\hat{u}}$ and $w_{2,t,\hat{u}}$ represent the first and the second best hypotheses respectively in utterance $\hat{u}$ and position t, α and β are the hyperparameters, $\mathbf{kld}()$ represents the function to calculate Kullback-Leibler distance and $\mathbf{I}()$ is a binary indicator function which return 1 only when the conditions are satisfied.

In equation 3, the local-global uncertainty of the ASR hypothesis is the weighted local uncertainty. The weight determined by the positive and negative samples observed in the data pool. The positive samples reduce the uncertainty and the magnitude of the reduction depends on the Kullback-Leibler distance between the posterior distribution of the word hypotheses in the current sample and those in the observed positive samples. Meanwhile, the negative samples increase the uncertainty of the current sample in the same manner.

The uncertainty based SSL data selection can be performed by selecting the utterances with the low uncertainties. While the uncertainty can be calculated either using local uncertainty only, or using both local and global uncertainties.

Compared to the GERM algorithm in [9], this work dropped the process of adjusting the utterance entropies during data selection. This significantly accelerated the data selection process. However, the disadvantage is that the redundant training samples may be selected. To reduce the redundancy, this work used a simple method. The utterances with low uncertainty can be divided into two parts. In the first part of data, the local uncertainty of the utterance was already low. In the second part of data, the local uncertainty of the utterance was

not low, but after introducing the global uncertainty, the utterance uncertainty was reduced. Considering that the utterances with low local uncertainty usually have already been recognized well by the current system and do not contain too much new information, the random sampling was performed to this part of data to reduce redundancy. While for the second part of data, since they got higher local uncertainty, it is likely that they contained the new information for the ASR system to learn. Thus the random sampling was not performed.

2.3. Margin based SSL

Margin based methods have been used in AL very successfully [14] [15]. This work introduced the margin based data selection into SSL. The idea of the margin based data selection is that the training samples close to the decision boundary should be important to the performance of the recognition system, thus should be selected, as shown in (a) of figure 2. However, in margin based methods, the calculation of the distance between the sample and the decision boundary is required. For DNN based ASR system, this calculation is non-trivial. This work considered the margin based methods in a different way. Since directly finding the samples close to the boundary is not easy, this work tried to achieve it in an indirect way, as shown in (b) of figure 2.

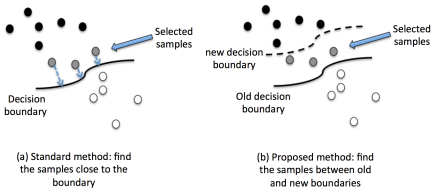

Figure 2: *Margin based data selection*

Instead of finding the samples close to the boundary directly, the proposed method moves the decision boundary. This way, all the samples fall in the area between the old boundary and the new boundary should be selected. More concretely, this work moved the decision boundary by tuning the LM. A strong LM with in-domain data and a weak LM without in-domain data were trained. Then based on the same AM, the uncertainties of the utterances in the data pool were calculated separately with two different LMs. The selected utterances can be expressed as:

$$\mathsf{S} = \left\{u|\mathbf{UNC}(u|\mathsf{LM}_{\mathsf{weak}}) > t_1\right\} \bigcap \left\{u|\mathbf{UNC}(u|\mathsf{LM}_{\mathsf{strong}}) < t_2\right\} \quad (4)$$

where $\mathbf{UNC}(u|\mathsf{LM}_{\mathsf{weak}})$ and $\mathbf{UNC}(u|\mathsf{LM}_{\mathsf{strong}})$ are the uncertainty of the utterance u based on the weak and strong LM respectively. t_1 and t_2 are the hyperparameters. To make the margin based selection meaningful, t_1 should be bigger than t_2. That says, the selected utterances should have lower uncertainty with strong LM and higher uncertainty with weak LM.

3. Experiments

The experiments in this work were based on a far-field ASR system in Japanese language. The AM is a LSTM based model. Transfer learning [16] was used in the AM training. Both the cross-entropy training and the bMMI sequence training were used to train the AMs.

The SSL experiments were performed in two different scenarios. The first one is that the in-domain training data with human annotation was not available. Thus the initial AM was trained by the out-domain data which mismatches to the untranscribed data in acoustic and linguistic characteristics. In the second scenario, small proportion of the in-domain data (less than 3%) had been transcribed by human, and it was used to train the initial AM. The initial AM trained with the in-domain data outperformed the AM trained with the out-domain data by about 20% relatively. The overall machine transcription quality was greatly improved with a small proportion of the human transcribed in-domain data. In this work, the data selections were performed separately based on these two machine transcriptions with different qualities.

3.1. SSL initialized with out-domain data

Using the out-domain AM and the trigram LM that excluded the in-domain data, the training speech corpus was automatically transcribed for SSL.

Several SSL training sets were generated based on different selection strategies. The first training set was generated with local uncertainty. Since the local uncertainty based SSL only considers the information of the current utterance, it is quite close to the confidence based SSL. The second training set was generated by considering both local and global uncertainty. Compared to the local uncertainty based SSL, the second training set introduced about 10% more training samples which was contributed from the global uncertainty. As mentioned in section 2.2, to reduce the redundancy this work also introduced the random sampling to the utterances with low local uncertainty. This yielded the third training set. The three SSL training sets being investigated in this experiment can be summarized as follows:

- local: local uncertainty based SSL
- local+global: local-global uncertainty based SSL
- rs_local+global: local-global uncertainty based SSL, random sampling the data with low local uncertainty

In the experiments with out-domain data initialization, the margin based SSL was not performed. These experiments assumed that the in-domain data was not available. Thus the strong LM which included the in-domain data could not be trained. Although there are other ways to tune the LMs, e.g. LM pruning etc., they were not investigated in this work. The ASR performance was given in Table 1.

The SSL data selection experiments indicate that combining the local and global uncertainty significantly improve the ASR performance. Although only about 10% additional training samples were introduced by global uncertainty information, it significantly improved the ASR performance of SSL. Meanwhile, random sampling the data with low local uncertainty is critical to improve the performance of the SSL. It reduced the size of the selected data from $4.9M$ utterances to $1.1M$, but the ASR performance was improved by 7% relatively. This can be explained by the fact that the useful information from the data with low local uncertainty is very limited. Large amount of this kind of data did not improve the ASR performance, but weakened the contribution from the important data. In total, the SSL methods in this work improved the ASR performance by 17% relatively when only the out-domain human annotated training data was available.

Another interesting result in Table 1 is the result of bMMI training. Based on SSL data selection proposed in this work, the sequence training improved the ASR performance compared to the cross-entropy training. While in the previous work of SSL

Table 1: *Word error rate reduction results for SSL initialized with out-domain data*

data selection method	# selected utterances	Training method	WERR [%] Dev	WERR [%] Test
initial model without SSL		bMMI	0.0	0.0
local	4.5M	XE	6.93	8.78
local+global	4.9M	XE	10.86	10.56
rs_local+global	1.1M	XE	14.50	14.49
rs_local+global	1.1M	bMMI	17.29	17.34

Table 2: *Word error rate reduction results for SSL initialized with in-domain data*

data selection method	# selected utterances	Training method	WERR [%] Dev	WERR [%] Test
supervised (initial model without SSL)		bMMI	0.0	0.0
rs_local+global	1.46M	bMMI	0.57	0.85
margin	0.57M	bMMI	-6.55	-6.03
margin+supervised	0.57M	bMMI	0.85	1.32
rs_local+global+margin+supervised	1.63M	bMMI	3.51	3.70

e.g. [6], [17], no gain was observed in sequence training when the automatically transcribed data was used. This result indicates that the SSL method proposed in this work did catch some important training samples compared to the traditional confidence based methods.

3.2. SSL initialized with in-domain data

In this experiment, the SSL was initialized with an AM trained by a small set of the in-domain data which is homogeneous to the data in data pool. Thus the quality of the machine transcriptions in this experiment is much better than the one in subsection 3.1. At first, the data selection based on local-global uncertainty was investigated with the same strategy as "rs_local+global" in subsection 3.1. Then, the margin based method was investigated by two models. One was trained only using the automatically transcribed data from the margin based data selection. The other was trained by merging the margin based selected data and the supervised data used for initial AM training. Finally, the supervised data which was used to train the initial model was merged with all the automatically transcribed data from SSL data selections. The experimental results were shown in Table 2.

Using the in-domain data to initialize SSL, the performance of the initial model, i.e. the baseline was significantly improved. The gain from SSL is not as big as the results based on the out-domain data initialization. Simply using automatically transcribed data from "rs_local+global" method, the AM achieved comparable performance to the initial model trained by the supervised data. In the experiments of margin based method, the model trained by the automatically transcribed data alone did not achieve the good performance. Only when the supervised data was merged with the automatically transcribed data, the gain was observed. This is explainable since the margin based data selection relies on the decision boundary which was defined by the initial AM. In other words, the role of the selected data from margin based method is the complement of the initial model. Thus it should be used together with the initial supervised data. Finally, merging all the automatically transcribed data with the initial data, about 3.7% relative WERR was observed.

Compared to the experiments with out-domain data initialization, the gain from SSL was much smaller when in-domain data was used for initialization. This may be explained by the fact that the in-domain data is much more efficient to improve the ASR performance. The in-domain initial model outperformed the out-domain initial model by about 20% relatively. It was good enough to cover most of the information that can be learned from the SSL, thus reduced to room of the improvement.

4. Conclusions

This work investigated the SSL methods to bootstrap the AM training. Two SSL methods were proposed in this work. The local-global uncertainty based method selects the speech utterances not only based on the local uncertainty of the current utterance, but also the global uncertainty learned from the whole data pool. The margin based method selects the utterances which are close to the decision boundary and the data selected was implemented by adjusting the decision boundary rather than calculating the distance between the samples and the boundary directly. Using the AM trained by out-domain data as baseline and initial model, the proposed method can achieve about 17% WERR without any extra human annotated data. While using the in-domain data to build the initial model, 3.7% WERR was observed. Another interesting result in this work is that the ASR gain was observed by bMMI sequence training. While from the previous work based on confidence model, the sequence training did not improve the ASR due to the imperfectness of the automatically transcribed data.

5. Acknowledgements

The author would like to thank Frederick Weber and Daniel Willett for the valuable discussions on the work of this paper.

6. References

[1] L. Lamel, J. L. Gauvain, and G. Adda, *"Unsupervised acoustic model training,"* in Proc. of ICASSP, 2002.

[2] Y. Huang, Y. Wang, and Y. Gong *"Semi-supervised training in deep learning acoustic model."* in Proc. of Interspeech, 2016.

[3] Liao, H., McDermott, E., and Senior A. *"Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription,"* in Proceeding of IEEE ASRU 2013.

[4] Yu, K., Gales, M., Wang, L., and Woodland, P. C., *"Unsupervised training and directed manual transcription for LVCSR"*, Speech Communication, vol. 52, 2010.

[5] Thomas Kemp, Alex Waibel *"Unsupervised Training of a Speech Recognizer: Recent Experiments,"* Proc. of EUROSPEECH, 1999.

[6] T. Drugman, J. Pylkkonen nad R Kneser *"Active and semi-supervised learning in asr: Benefits on the acoustic and language models,"* Proc. of Interspeech 2016

[7] Lori Lamel, Jean-Luc Gauvain and GillesAdda *"Lightly supervised and unsupervised acoustic model training,"* Computer Speech and Language, Jan. 2002

[8] Sheng Li, Yuya Akita and Tatsuya Kawahara *"Semi-Supervised Acoustic Model Training by Discriminative Data Selection From Multiple ASR Systems' Hypothes,"* IEEE/ACM Transactions on Audio, Speech, and Language Processing,Volume 24,Issue 9 ,Sept. 2016

[9] Dong Yu, Balakrishnan Varadarajan, Li Deng and Alex 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, no. 3, 2010.

[10] Jinyu Li, Rui Zhao, Jui-Ting Huang and Yifan Gong *"Learning Small-Size DNN with Output-Distribution-Based Criteria,"* Proc. of Interspeech, 2014.

[11] Sree Hari Krishnan Parthasarathi, Nikko Strom *"Lessons from building acoustic models from a million hours of speech,"* Proc. of ICASSP, 2019.

[12] Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian and Kai Yu *"Knowledge Distillation for Sequence Model,"* Proc. of Interspeech, 2018.

[13] 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, Oct. 2000

[14] Maria-Florina Balcan, Andrei Broder and Tong Zhang *"Margin Based Active Learning,"* International Conference on Computational Learning Theory, 2007

[15] Melanie Ducoffe and Frederic Precioso *"Adversarial Active Learning for Deep Networks: a Margin Based Approach,"* `https://arxiv.org/pdf/1802.09841.pdf`

[16] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng and Yifan Gong *"Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,"* in proc. of ICASSP 2013

[17] Karel Veselý, Mirko Hannemann and Lukáš Burget *"Semi-supervised training of Deep Neural Networks,"* in proc. of ASRU 2013