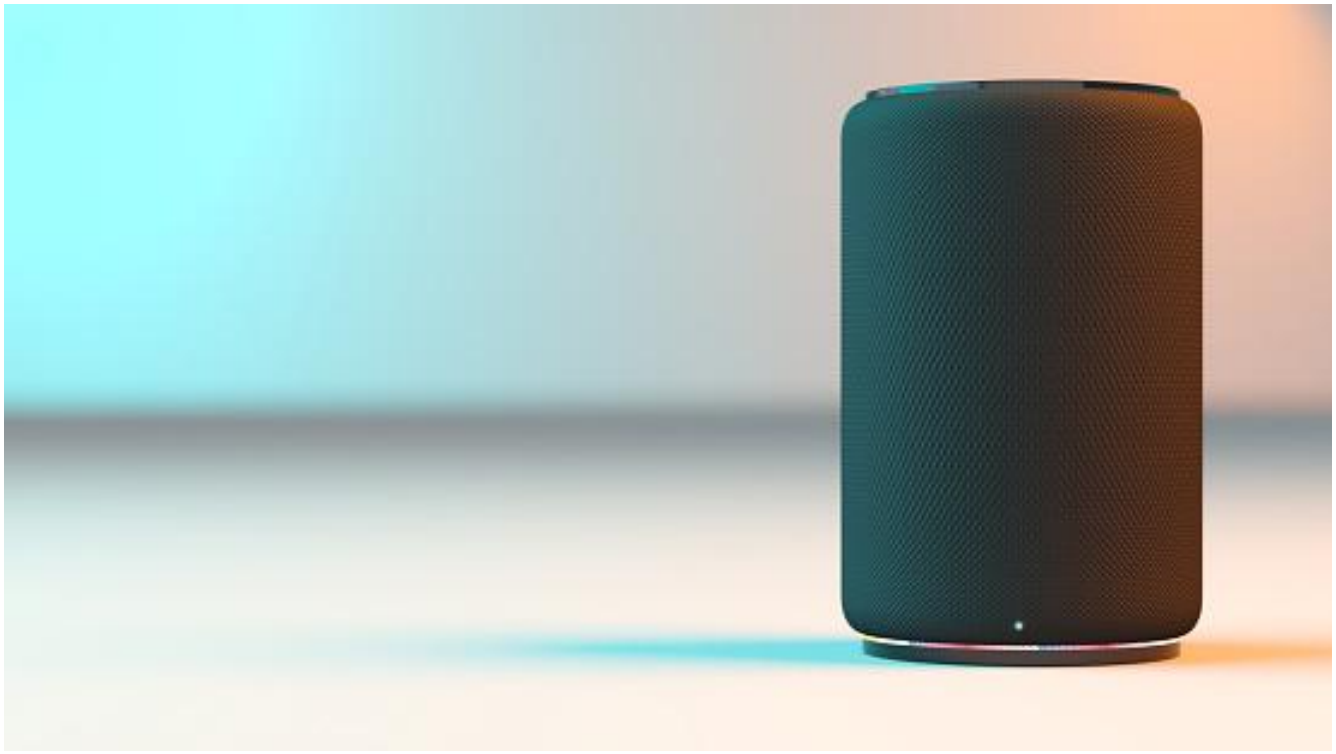
EXPERIMENT 10



ALEXA

Group Members :

Priyanshu Singh -53
Arnav Singhal -56
Subrato Tapaswi -60

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 a small amount of the in-domain data. Index Terms: speech recognition, semi-supervised training.

INTRODUCTION :

Bootstrapping the acoustic model (AM) training for the automatic speech recognition (ASR) system building with a 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 has become an important research area since nowadays large amounts 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 method was 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 deep learning dominates the research in speech domain, 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 GERMmethodproposed 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 a confusion network. Meanwhile, this work kept the idea of GERMto 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 the original GERM method were 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 SQLmethods 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. 2.3. 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 systems, 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, 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 that 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

LMwithin-domaindataandaweakLMwithoutin-domaindata 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:
S =u|UNC(u|LMweak) > t1 u|UNC(u|LMstrong) < t2 (4) where UNC(u|LMweak) and UNC(u|LMstrong)
are the uncertainty of the utterance u based on the weak and strong LM respectively. t1 and t2 are the hyperparameters. To make the margin based selection meaningful, t1 should be bigger than t2. That says, the selected utterances should have lower uncertainty with strong LM and higher uncertainty with weak LM.

Conclusion :

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.