An Active Learning Approach to Task Adaptation

Ji Wu[1], *Zhiyang He*[2], *Ping Lv*[2]

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Tsinghua-iFlytek Joint Laboratory for Speech Technologies, Beijing, China
wuji_ee@tsinghua.edu.cn, {zyhe_ts, luping_ts}@mail.tsinghua.edu.cn

Abstract

An active learning approach is proposed to automatically analyze speech recognition tasks and select particularly useful adaptation data. In this approach, the distribution of task data is first estimated, which is a combination of two distributions based on N-best recognition results and low confidence data. After that, a subset of adaptation data is selected in two stages using a greedy algorithm according to the estimated distribution. Low confidence data are firstly selected and manually labeled. Then, the high confidence data are selected based on the top-best recognition results, which are also used as labels for the adaptation. The experimental results of the subsequent task adaptation show that the proposed active learning approach can effectively select the useful data to improve the overall performance of the system. The word accuracy is close to, and even exceed, the performance of supervised adaptation using all of the data, when only 10%-20% of the total data need to be manually labeled.

Index Terms: task adaptation, task analysis, data selection, active learning, speech recognition

1. Introduction

The performance of task adaptation highly depends on the amount of labeled task-specific adaptation data. It might be easy to collect a mass of task-specific data for many practical applications. However, data labeling is always difficult.

In [1][2][3][4], unsupervised adaptation approach was proposed by using automatically recognized labels. However, no significant performance improvements are derived in many realistic applications due to lots of wrong labels. In [5], a subset of training data was selected with high confidence score so that the automatic labels of selected data were more reliable. But for one certain task, the relevant speech data, which is of low confidence and is transcribed manually afterwards, is usually more useful. In [6][7], algorithms of selecting a set of adaptation or training data according to the task-specific data distribution were proposed. It was reasonable to actively select data with a good representation to a specific task. However, it is always difficult to estimate a good distribution of the task because of the lack of true labels. In [8][9][10][11][12], the idea of active learning, studied extensively in the field of machine learning, was applied to several spoken language processing applications. In their studies, a subset of training data with high uncertainties was selected to be manually labeled, which was considered most helpful for the learning purpose. However, the distribution of the data with low confidence could be different from the one of the whole data and using these data cannot guarantee the performance improvement.

This paper presents an automatic approach to analyze a given task in order to obtain a distribution related to the task and select adaptation data based on the obtained distribution, concerning both uncertainties and representations. The rest of the paper is organized as follows. Section 2 describes how to estimate the task distribution. Section 3 describes a two-stage greedy procedure of data selection. Experimental results are presented in Section 4. Conclusions are given in Section 5.

2. Task Distribution

2.1. Overview

A voice search application is used as the ASR task for investigation of the proposed active learning approach. In this application, users can speak the *tag-names* within a certain inventory, e.g. names/titles of singers/songs. All of the *tag-names* form the task vocabulary. Table 1 shows a voice search corpus, which is also used in our experiments. The corpus includes two sets of acoustic data denoted *S04* and *S05*. The corresponding vocabulary in the corpus consists of 4282 different *tag-names*.

Table 1: *Voice search speech corpus.*

Corpus	Recording Time	Utterance Number
S04	April-2008	10114
S05	May-2008	8532

For many actual ASR systems, the application information and conditions can be considered slowly time-variant. So we can use the current data to perform adaptation, estimate the data distribution, etc. In our experiments, data in set *S04* are firstly recognized by a task independent (TI) ASR system for obtaining the N-best results which are then used for estimating the distribution of the *tag-names*. Set *S04* is also used for the subsequent data selection. Set *S05* is used as a test set.

2.2. Occurrence Distribution Estimation

Firstly, a simple distribution can be estimated according to N-best recognition results. This distribution approximatively reflects the occurrence probabilities of the *tag-names* in the task. Denote the set of candidate utterances by $U = \{u_i; i = 1, \cdots, |U|\}$ and $|\cdot|$ refers to the size of a set, that is, $|U|$ is the number of candidate utterances in U. Denote by $W = \{w_j; j = 1, 2, \cdots, |W|\}$ the set of *tag-names* in the task vocabulary. Given an utterance u_i, the posterior probability [13] of *tag-name* w_j, which is dependent on N-best results, is denoted by $PP(w_j|u_i)$. For each *tag-name* $w_j \in W$, the *N-best occurrence distribution* is estimated as

$$P_{Occ}^{TD}(w_j) = \frac{\sum_{i=1}^{|U|} PP(w_j|u_i)}{|U|} \quad (1)$$

where *TD* denotes "task distribution".

2.3. Confusability Distribution Estimation

In order to reflect and emphasize the existence of the *tag-names* with higher confusability in the vocabulary, an N-best confusability distribution is estimated by using only the utterances with low confidence scores.

A confidence score depended on the posterior probability of each utterance u_i is defined as $CM(u_i) = PP(\hat{w}_1(u_i)|u_i)$, where $\hat{w}_1(u_i)$ is the first one of the N-best recognition results of utterance u_i and $\hat{w}_1(u_i) \in W$. $CM(\cdot)$ lies in the $[0, 1]$ interval. For each *tag-name* $w_j \in W$, the *N-best confusability distribution* is computed as

$$P_{Con}^{TD}(w_j) = \frac{\sum_{i=1}^{|U|} \left(PP(w_j|u_i)\delta\big(CM(u_i) \le TH_c\big)\right)}{\sum_{i=1}^{|U|} \delta\big(CM(u_i) \le TH_c\big)} \quad (2)$$

where $\delta(\cdot)$ is the indicator function, that is, $\delta(\pi)$ is 1 when the π is true and 0 otherwise. TH_c is a predefined threshold. Only those utterances u_i, the confidence score of which is less than TH_c, are used in confusability distribution estimation.

2.4. Combination

The equation (1) and equation (2) above depict the statistical distribution of the *tag-names* in the task from different perspective. The former one reflects the probabilities of whole data, while the latter one focuses on the low confidence data which is more useful for improving the performance. A weighted sum strategy can be adopted to combine the two probabilities calculated in section 2.2 and section 2.3. For each *tag-name* w_j, the final distribution is calculated as

$$P^{TD}(w_j) = \alpha P_{Occ}^{TD}(w_j) + (1-\alpha)P_{Con}^{TD}(w_j) \quad (3)$$

where $0 < \alpha < 1$ is the weight to balance the effects between P_{Occ}^{TD} and P_{Con}^{TD}. Therefore, the probabilities of almost all *tag-names* contribute to the task data distribution and the task data distribution is then used to form the objective functions for adaptation utterance selection.

3. Two-Stage Data Selection

3.1. Overview of the Data Selection

The Kullback-Leibler(KL) measure[14], which is also known as the cross-entropy, is widely-used when the "closeness" between two probability distributions need to be calculated. The value of KL measure is greater than or equal to 0 and only when one distribution is equivalent to the other one, can the value 0 be reached. So selecting a data set, the distribution of which is "closest" to a given distribution, is generally formulated as a discrete optimization problem [15][16].In this paper, a greedy algorithm[6] based on KL measure is used. In our task, the KL measure in discrete case is defined as:

$$I(P^{S}||P^{TD}) = \sum_{j=1}^{|W|} P(w_j;S)\log\frac{P(w_j;S)}{P^{TD}(w_j)} \quad (4)$$

where $S = \{\hat{u}_k; 1 \le k \le |S|\}$ denotes the sets of selected utterances and $\hat{u}_k \in U$, $P(w_j;S)$ denotes the probability of w_j in S and $P^{S} = \{P(w_j;S); w_j \in W, j = 1,2,\ldots,|W|\}$, P^{TD} is the estimated task distribution and is calculated using the equation (3).

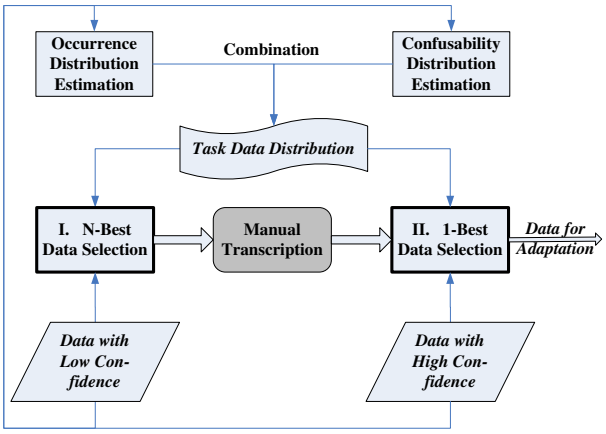

Figure 1: *Outline of Task Analysis and Data Selection.*

Fig.1 illustrates the procedure of task analysis and data selection. Data selection will be conducted after obtaining the distribution of the task. In our experiments, the utterances in data set *S04* are used as candidates on the two-stage utterance selection. All of the utterances selected in the two stages are used for the model adaptation to improve the performance of the ASR system. The detail of the data selection procedure is illustrated in the next two sections.

3.2. Data Selection from Low Confidence Candidates

On the first stage, an N-best data selection strategy is used to select the utterances from the candidates with low confidence score until a certain condition is met.

Let's denote by $U^{LC} = \{u_i^{LC}; 1 \le i \le |U^{LC}|\}$ a set of candidate utterances with low confidence score. Suppose we need to select N^{LC} utterances from U^{LC}. The procedure of the first stage utterance selection is shown in Fig.2.

1) $S_0 = \phi$

2) For i = 1 to N^{LC}

(a) If all $u^{LC} \in U^{LC}$ have been examined,
break;
Else
selecting the unexamined utterance u_{min}^{LC} such that

$$u_{min}^{LC} = \arg\min_{u^{LC} \in U^{LC}} I\left(P^{S_i^*} \| P^{TD}\right) \text{ with}$$

$$S_i^* = S_i \bigcup \left\{u^{LC}\right\}$$

$$P\left(w_j; S_i^*\right) = \frac{\sum_{i=1}^{|S_i^*|} PP\left(w_j \mid u_i\right)}{\left|S_i^*\right|}$$

$$P^{S_i^*} = \left\{P\left(w_j; S_i^*\right); w_j \in W, j = 1,2,...,|W|\right\}$$

(b) $S_{i+1} = S_i \bigcup \left\{u_{\min}^{LC}\right\}$, $U^{LC} = U^{LC} - \left\{u^{LC}\right\}$

3) $S^{LC} = S_i$

Figure 2: *First stage data selection algorithm.*

Once the first stage data selection is finished, the selected

utterances set denoted by S^{LC} are *transcribed manually* and the real count of each *tag-name* is derived, so $P(w_j; S^{LC})$ is changed based on the true occurrence. Then S^{LC} will be used as initial set for the next stage of data selection.

3.3. Data Selection from High Confidence Candidates

The process of the second stage data selection is similar to the first stage, except selecting data from the candidates with high confidence score and using the top-best recognition results. The primary purpose of data selection on this stage is to add data to S^{LC} so that the distribution of final selected data set is further close to the distribution P^{TD}.

1) $S_0 = S^{LC}$

2) For i = 1 to N^{HC}

(a) If all $u^{HC} \in U^{HC}$ have been examined, break;
Else
selecting the unexamined utterance $u_{\min}^{HC}$ such that

$$u_{\min}^{HC} = \arg \min_{u^{HC} \in U^{HC}} I\left(P^{S_i^*} \parallel P^{TD}\right) \text{ with}$$

$$S_i^* = S_i \bigcup \left\{u^{HC}\right\}$$

$$P\left(w_j; S_i^*\right) = \frac{N\left(w_j\right) + \sum_{i=1}^{|S_i^*|} \delta\left(W_1\left(u_i\right) = w_j\right)}{\sum_{l=1}^{|W|}\left[N\left(w_l\right) + \sum_{i=1}^{|S_i^*|} \delta\left(W_1\left(u_i\right) = w_l\right)\right]}$$

$$P^{S_i^*} = \left\{P\left(w_j; S_i^*\right); w_j \in W, j = 1, 2, ..., |W|\right\}$$

(b) $S_{i+1} = S_i \bigcup \left\{u_{\min}^{HC}\right\}$, $U^{HC} = U^{HC} - \left\{u^{HC}\right\}$

3) $S^{LHC} = S_i$

Figure 3: *Second stage data selection algorithm.*

We denote by $N(w_j)$ the number of selected utterances belonging to *tag-name* w_j on the first stage. The utterance selection of second stage is then processed as shown in Fig.3. The final selected data set is denoted by S^{LHC}, which is used for the subsequent task adaptation.

4. Experiments and Results

4.1. System Setups and Evaluation

The basic speech units of our ASR system are right context dependent initial and toned-final (RCI/TF). There are 253 basic units in total. Each right-context-initial basic unit is modeled by a three-emitting-state left to right continuous density hidden Markov model (LR-CDHMM) without state skipping and the toned-final unit is modeled by a five-emitting-state LR-CDHMM without state skipping. Each state has 12 Gaussian mixture components with each component having a diagonal covariance matrix. A special three-state CDHMM is also employed for silence modeling. The 43-dimensional feature vector used in the system includes 12-dimension MFCC, log-scaled energy normalized by the average energy of the individual sentence, and their first and second order derivatives. Sentence-based cepstral mean subtraction is applied for acoustic normalization both in training and test. Moreover there are 4-dimensional tone features. HTK toolkit [17] is used to train a set of speaker-independent RCI/TF model with 1065 states in total. The training data consists of about 142-hours speech data including 1,106k Chinese syllables. Given the above set of TI CDHMMs, a baseline word accuracy of 69.9% can be achieved on set *S05*.

The adaptation algorithm in all experiments uses a two-pass strategy. In the first pass, a standard MLLR adaptation[1] with global transformation is adopted and the adapted model is used as the seed model in the next pass. In the second adaptation step, the MLLR adaptation with multiple regression classes and the MAP adaptation[18] are combined to improve the performance[17]. In all experiments, the regression class tree of MLLR has 149 leaf nodes and is built from the TI CDHMMs. The weighting of the a priori knowledge in MAP adaptation is 12.0.

In the experiments, we use 3-best recognition results of each utterance in set *S04* during the task analysis and the data selection. We choose 0.9 as the threshold(TH_c) for splitting the set *S04* into two parts, high confidence data and low confidence data. The weight α for combination two distributions in section 2.4 is set to 0.2.

Four different data selection approaches below are conducted for comparing with our approach. The data selected in each approach is used for model adaptation based on TI CDHMMs.

Supervised Approach: Total 10114 utterances in set *S04* are used as adaptation data and the manual labels are used.

High Confidence Unsupervised Approach: Only the utterances with confidence score higher than $TH_c = 0.9$ in *S04* are used as adaptation data and the 1-best recognition results are used as the corresponding labels[5]. In our experiments, totally 3594 utterances are selected for subsequent adaptation.

Random Approach: A certain amount of utterances with low confidence scores less than $TH_c = 0.9$ in *S04* are randomly selected and labeled manually. Then, all of the utterances with high confidence score (3594 utterances) in set *S04* are added into the selected data set, and the 1-best recognition results are used as their labels.

Lowest Confidence Active Approach: A certain amount of utterances with the lowest confidence scores in *S04* are selected and labeled manually[11]. All of the utterances with high confidence score are then added into the selected data set as well as the *random approach*.

4.2. Experimental Results

Fig.4 shows the experimental results of the four aforementioned approaches in section 4.1 and our proposed approach. The supervised adaptation system with 79.8% word accuracy is obtained in our experiment, while performance of the high confidence unsupervised adaptation is 73.4%. Three numbers, 500, 1000 and 2000, are chosen as the upper bound to control the amount of the manually labeled data in the *random approach*, *lowest confidence active approach* and our approach. For the *random approach*, three groups of experiments corresponding to the numbers of the manually transcribed data were conducted. For each experimental group, we repeated the experiment five times because of the random selection. Table 2 shows the performances and Fig.4 gives the average accuracies.

The performance of the *lowest confidence active approach* is similar to the *random approach*'s. For the *lowest confidence active approach*, the word accuracies of adaptation systems with different size of manually labeled data are 75.8%(500), 77.7%(1000), 78.9%(2000). The corresponding performance of

Table 2: *Performance of the Random Approach.*

manually labeled	word accuracy(%)					average (%)
	1	2	3	4	5	
500	75.71	75.62	75.57	75.53	75.71	75.63
1000	77.47	77.14	77.30	77.56	77.11	77.32
2000	79.31	79.14	79.11	79.19	79.38	79.23

our approach are 79.0%, 79.6%, 80.4%, which are 10% better on average than the *random approach* and the *lowest confidence active approach*. It can be observed as well that, the performance of the proposed approach with 500 utterances labeled manually exceeds the *lowest confidence active approach*'s, in which 2000 utterances are labeled manually. When 1000(about 10% of all) utterances are labeled in our approach, the performance is close to the supervised approach using the whole data set. Moreover, when the number of manually labeled data reaches 2000(about 20% of all), the performance exceeds the one of the supervised approach.

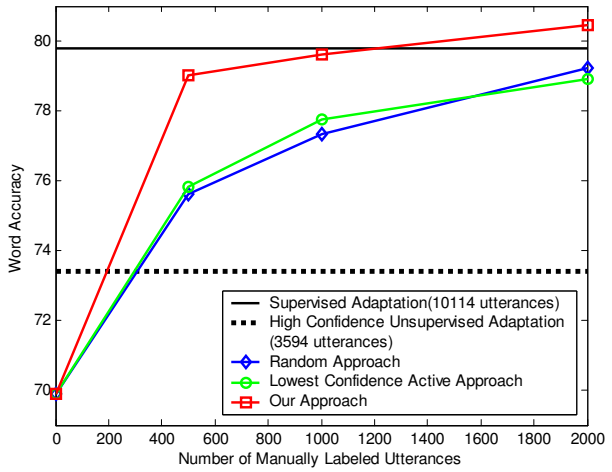

Figure 4: *Experimental results of the five approach.*

Experimental results demonstrate that, for a speech recognition task, the distribution of the low confidence data which is always not correctly recognized is more important, not only for estimating the distribution of the task but also for the data selection. Selecting a small amount of these data according to the estimated distribution and transcribing them manually are greatly helpful to performance improving. Meanwhile, the distribution of the total data, containing the high confidence data, is also necessary to be taken into account. Because only using the data with low confidence, even when the data are transcribed manually, may cause a big difference between the distribution of selected data and the one of the whole task. It is necessary to add the data with high confidence into the selected data set, so as to make the distribution more consistent with the task distribution. Therefore, during the estimation of the task distribution and the data selection, it is reasonable to utilize the information of both the low and high confidence data.

5. Conclusions

This paper presents an active approach to effectively select adaptation data according to the estimated task distribution. Both the uncertainties of the low confidence data and the representations of the whole data are considered in all of the procedures. Experimental results show that the proposed active learning approach performs well when only labeling a fraction of total data.

We are currently investigating an extension of this work to more recognition tasks, such as LVCSR. We are also looking into the more quantitative relationship between low confidence data and the performance for future work.

6. References

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, pp.171-185, 1995.

[2] P. C. Woodland, D. Pye and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression", *Proc. ICSLP*, 1996, pp.1133-1136.

[3] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[4] L. Lamel, J. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 877–880.

[5] T. Anastasakos and S. V. Balakrishnan, "The use of confidence measures in unsupervised adaptation of speech recognizers," in *Proc. ICSLP*, 1998.

[6] X. Cui and A. Alwan, "Efficient adaptation text design based on the Kullback-Leibler measure," *Proc. ICASSP*, 2002, pp.I-613-616.

[7] J.-L. Shen, H.-M. Wang, R.-Y. Lyu and L.-S. Lee, "Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition," *Computer Speech and Language*, Vol. 13, pp.79-97, 1999.

[8] D. Hakkani-Tur, G. Riccardi, and G. Tur, "An active approach to spoken language processing," *ACM Trans. on Speech and Language Processing*, Vol. 3, No. 3, pp.1-31, 2006.

[9] Q. Huo and W. Li, "A DTW-based dissimilarity measure for left-to-right hidden Markov models and its application to word confusability analysis," *Proc. Interspeech - ICSLP*, 2006, pp.2338-2341.

[10] Q. Huo and W. Li, "An active approach to speaker and task adaptation based on automatic analysis of vocabulary confusability," in *Proc. Interspeech*, 2007, pp.1569-1572.

[11] T. M. Kamm and G. G. L. Meyer, "Robustness aspects of active learning for acoustic modeling," *Proc. ICSLP*, 2004, pp.1095-1098.

[12] H.-K. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding," *Proc. Interspeech - Eurospeech*, 2005, pp.437-440.

[13] B. Rueber, "Obtaining confidence measure from sentence probabilities," *Proc. Eurospeech*, 1997, pp. 739-742.

[14] Solomon Kullback, *Information theory and statistics*, John Wiley & Sons, 1959.

[15] H. Kawai, S. Yamamoto, N. Higuchi, and T. Shimizu, "A design method of speech corpus for text-to-speech synthesis taking account of prosody," *Proc. ICSLP*, 2000, pp.420-425.

[16] J. P. H. van Santen and A. L. Buchsbaum, "Methods for optimal text selection," *Proc. Eurospeech*, 1997, pp.553-556.

[17] S. T. Young, et al., *The HTK Book*(revised for HTK Version 3.4), 2006.

[18] J. L. Gauvain and C. -H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp.291-298, 1994.