ACTIVE LEARNING FOR ACCENT ADAPTATION IN AUTOMATIC SPEECH RECOGNITION

Udhyakumar Nallasamy[1]*, Florian Metze*[1] *and Tanja Schultz*[1,2]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Cognitive Systems Labs, Karlsruhe Institute of Technology, Karlsruhe, Germany
{unallasa, fmetze, tanja}@cs.cmu.edu

ABSTRACT

We introduce a novel active learning algorithm for speech recognition in the context of accent adaptation. We adapt a source recognizer on the target accent by selecting a matched subset of utterances from a large, untranscribed and multi-accented corpus for human transcription. Traditionally, active learning in speech recognition has relied on uncertainty based sampling to choose the most informative samples for manual labeling. Such an approach doesn't include explicit relevance criterion during data selection, which is crucial for choosing utterances to match the target accent, from datasets with wide-ranging speakers of different accents. We formulate a cross-entropy based relevance measure to complement uncertainty based sampling for active learning to aid accent adaptation. We evaluate the algorithm on two different setups for Arabic and English accents and show that our approach performs favorably to conventional data selection. We analyze the results to show the effectiveness of our approach in finding the most relevant subset of utterances for improving the speech recognizer on the target accent.

Index Terms— Automatic Speech Recognition, Accent Adaptation, Active Learning

1. INTRODUCTION

Speech interfaces are becoming pervasive among the common public with the prevalence of smart phones and cloud-based computing. This pushes the Automatic Speech Recognition (ASR) systems to cater to wide-ranging speakers with varying accents. Accent adaptation in ASR typically involves adapting an acoustic model trained on the source accent to the target accent, using relatively small amounts of adaptation data. It is prohibitively costly to obtain large accented speech datasets, due to the effort involved in collecting and transcribing speech, even for a few of the major accents. On the other hand, for tasks like Broadcast News (BN) or Voice search, it is easy to obtain large amounts of speech data with representative accents. However, such datasets seldom have accent markers or transcriptions. To make use of these speech collections, it is necessary to identify the most appropriate subset of data, when transcribed, provide the largest improvement in ASR accuracy for the specified target accent. Active learning algorithms can aid in finding such a relevant subset from vast amount of untranscribed data for human annotation, thereby greatly reducing the cost to build accurate, accent-specific ASRs.

Active learning is a commonly used machine learning technique in fields where the cost of labeling the data is quite high [1]. It has been applied in natural language processing [2], spoken language understanding [3], speech recognition [4, 5, 6, 7], etc. Many of the approaches relied on some form of uncertainty based measure for data selection. The assumption is that adding the most uncertain utterances provide the maximum information for re-training the classifier in the next round. Confidence scores are typically used for active learning in speech recognition [8] to predict uncertainty. Lattice [6] and N-best [7] based techniques have been proposed to avoid outliers with 1-best hypothesis. Representative criterion in addition to uncertainty have also been shown to improve data selection in some cases [9, 7].

Most of these algorithms strive to find the smallest subset from the untranscribed data set, which when labeled and used to re-train the ASR will have the same effect of using the entire dataset for re-training, thereby reducing the cost. However, in the case of accent adaptation using a dataset with multiple accents, our goal is not to identify the representative subset but to choose relevant utterances that best match the target test set. In this paper, we introduce a relevance criterion in addition to uncertainty based informative measure for data selection to match the target accent. We start with the ASR trained on a source accent. We use a relatively small, manually labeled adaptation data to adapt the recognizer to the target accent. We employ the adapted model to choose utterances from a large, untranscribed mixed dataset for human transcription, to further improve the performance on the target accent. To this end, we calculate cross-entropy based measure based on adapted and unadapted model likelihoods, to assess the relevance of an utterance. We combine this measure with uncertainty based sampling to choose an appropriate subset for manual labeling. We evaluate our technique on

Arabic and English accents and we achieve 50-87.5% data reduction for the same accuracy of the recognizer using purely uncertainty based data selection.

2. UNCERTAINTY BASED INFORMATIVENESS CRITERION

In speech recognition, uncertainty is quantified by the ASR confidence score. It is calculated from the word-level posteriors obtained by consensus network decoding [10]. As mentioned before, confidence scores calculated on 1-best hypothesis are sensitive to outliers and noisy utterances. [6] proposed lattice-entropy based measure and selecting utterances based on global entropy reduction. [7] observed that lattice-entropy is correlated with the utterance length and showed N-best entropy to be an empirically better criterion. In this work, we also use a entropy-based measure as informative criterion for data selection. We calculate the average entropy of the alignments in the confusion network as a measure of uncertainty of the utterance with respect to the ASR. It is given by

$$Informative\ score\ u_i = \frac{\sum_{\mathcal{A}\in u} E_{\mathcal{A}} T_{\mathcal{A}}}{\sum_{\mathcal{A}\in u} T_{\mathcal{A}}} \tag{1}$$

where $E_{\mathcal{A}}$ is the entropy of an alignment A in the confusion network and $T_{\mathcal{A}}$ is the duration of the link with best posterior in the alignment. $E_{\mathcal{A}}$ is calculated over all the links in the alignment.

$$E_{\mathcal{A}} = -\sum_{\mathcal{W}\in A} P_{\mathcal{W}} log\ P_{\mathcal{W}} \tag{2}$$

3. CROSS-ENTROPY BASED RELEVANCE CRITERION

In this section, we derive cross-entropy based relevance criteria for choosing relevant target accent utterances from the mixed set for human annotation. We formulate the source-target mismatch as a sample selection bias problem [11, 12, 13] under two different setups. In the first case, the source data consists mixed set of accents and the goal is to adapt the model trained on the source data to the specified target accent. The source model has seen the target accent during training, albeit it is under-represented along with other accents in the source data. In the second case, the source and target data consist of dissimilar accents and the source model is adapted to an unseen target accent. The following sections elaborate the derivation for each case.

3.1. Multi-accented case

In this setup, the source data contains a mixed set of accents. The target data, a subset of the source represents utterances that belong to a specific target accent. An utterance u in the data set is represented by a sequence of observation vectors and its corresponding label sequence. Let X denote the space of observation sequences and Y the space of label sequences. Let S denote the distribution over utterances $U \in X \times Y$ from which source data points (utterances) are drawn. Let T denote the target set distribution over $X \times Y$ with utterances $\hat{U} \subseteq U$. Now, utterances in T are drawn by biased sampling from S denoted by the random variable $\sigma \in \{0, 1\}$ or the *bias*. When $\sigma = 1$, the randomly sampled $u \in U$ is included in the target dataset and when $\sigma = 0$ it is ignored. Our goal is to estimate the bias $Pr[\sigma = 1|u]$ given an utterance u, which is a measure for how likely is the utterance to be part of the target data. The probability of an utterance u under T can be expressed interms of S as

$$Pr_T[u] = Pr_S[u|\sigma = 1] \tag{3}$$

By Bayes rule,

$$Pr_S[u] = \frac{Pr_S[u|\sigma = 1]Pr[\sigma = 1]}{Pr[\sigma = 1|u]} = \frac{Pr[\sigma = 1]}{Pr[\sigma = 1|u]} Pr_T[u] \tag{4}$$

The bias for an utterance u is represented by $Pr[\sigma = 1|u]$

$$Pr[\sigma = 1|u] = \frac{Pr_T[u]}{Pr_S[u]} Pr[\sigma = 1] \tag{5}$$

The posterior $Pr[\sigma = 1|u]$ represents the probability that a randomly selected utterance $u \in U$ from the mixed set, belongs to the target accent. It can be used as a relevance score for identifying relevant target accent utterances in the mixed set. Since we are only comparing scores between utterances for data selection, $Pr[\sigma = 1]$ can be ignored in the above equation as it is independent of u. Further, we can approximate $Pr_S[u]$ and $Pr_T[u]$, by unadapted and adapted model likelihoods. Substituting and changing to log domain,

$$Relevance\ Score\ u_r \approx \log\ Pr[u|\lambda_T] - \log\ Pr[u|\lambda_S] \tag{6}$$

The utterances in the mixed set can have different durations, so we normalize the log-likelihoods to remove any correlation of the score with the duration. The length normalized log-likelihood is also the cross-entropy of the utterance given the model [14, 15] with sign reversed. The score that represents the relevance of the utterance to target dataset is given by

$$Relevance\ Score\ u_r = (-H_{\lambda_T}[u]) - (-H_{\lambda_S}[u]) \tag{7}$$

where

$$H_\lambda(u) = -\frac{1}{T_u} \sum_{t=1}^{T_u} \log\ p(u_t|\lambda) \tag{8}$$

is the average negative log-likelihood or the cross-entropy of u according to λ and T_u is the number of frames in utterance u.

3.2. Dissimilar accents case

In this case, source and target correspond to two different accents. let A denote distribution over observation and label

sequences $U \in X \times Y$. Let S and T be the source and target distributions over $X \times Y$ and subsets of A, $U_S, U_T \subseteq U$. The source and target utterances are drawn by biased sampling from A governed by the random variable $\sigma \in \{0, 1\}$. When the bias $\sigma = 1$, the sampled utterance u is included in the target dataset and $\sigma = 0$ it is included in the source dataset. The distributions S and T can be expressed in terms of A as

$$Pr_T[u] = Pr_A[u|\sigma = 1]; Pr_S[u] = Pr_A[u|\sigma = 0] \quad (9)$$

By Bayes rule,

$$Pr_A[u] = \frac{Pr[\sigma = 1]}{Pr[\sigma = 1|u]} Pr_T[u] = \frac{Pr[\sigma = 0]}{Pr[\sigma = 0|u]} Pr_S[u] \quad (10)$$

Equating LHS and RHS

$$\begin{aligned} \frac{Pr_S[u]}{Pr_T[u]} &= \frac{Pr[\sigma = 1]}{Pr[\sigma = 0]} \frac{Pr[\sigma = 0|u]}{Pr[\sigma = 1|u]} \\ &= \frac{Pr[\sigma = 1]}{Pr[\sigma = 0]} \left[\frac{1}{Pr[\sigma = 1|u]} - 1 \right] \end{aligned} \quad (11)$$

As in the previous case, we can ignore the constant terms that don't depend on u as we are only comparing the scores between utterances. The relevance score, which is an approximation of $Pr[\sigma = 1|u]$ is given by

$$Relevance\ score\ u_r \approx \frac{Pr_T[u]}{Pr_T[u] + Pr_S[u]} \quad (12)$$

Changing to log-domain,

$$\begin{aligned} Relevance\ score\ u_r &\approx \log\ Pr_T[u] \\ &\quad - \log\ (Pr_T[u] + Pr_S[u]) \\ &= \log\ Pr_T[u] \\ &\quad - \log\ \left(Pr_T[u] \left[1 + \frac{Pr_S[u]}{Pr_T[u]} \right] \right) \\ &= -\log\ \left(1 + \frac{Pr_S[u]}{Pr_T[u]} \right) \end{aligned} \quad (13)$$

log is a monotonous function, hence $\log(1+x) > \log(x)$ and since we are only comparing scores between utterances, we can replace $\log(1 + x)$ with $\log(x)$. The relevance score is then the same as the multi-accented case

$$\begin{aligned} Relevance\ Score\ u_r &\approx \log\ Pr_T[u] - \log\ Pr_S[u] \\ &\approx \log\ Pr[u|\lambda_T] - \log\ Pr[u|\lambda_S] \end{aligned}$$

Normalizing the score with to remove any correlation with utterance length,

$$Relevance\ Score\ u_r = (-H_{\lambda_T}[u]) - (-H_{\lambda_S}[u]) \quad (14)$$

4. SCORE COMBINATION

Our final data selection algorithm uses a combination of relevance and uncertainty scores for active learning. The difference in cross-entropy is used a measure of relevance and the average entropy based on confusion networks is used as a measure of uncertainty or informativeness. Both the scores are in log-scale and we use a simple weighted combination to combine both the scores [7]. The final score in given by

$$Final\ score\ u_F = u_r * \theta + u_i \quad (15)$$

The mixing weight, θ is tuned on the development set. The final algorithm for active learning that uses both the relevance and informativeness scores is given below.

Algorithm 1 Active learning using relevance and informativeness scores

Input: $\mathcal{X}_T$:= Labeled Target Adaptation set ; $\mathcal{X}_M$:= Unlabeled Mixed set ; λ_S := Initial Model ; θ := Mixing weight $minScore$:= Selection Threshold

Output: λ_T := Target Model

$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
for all x in $\mathcal{X}_M$ **do**
  $Loglike_S := -CrossEntropy(\lambda_S, x)$
  $Loglike_T := -CrossEntropy(\lambda_T, x)$
  $Len := Length(x)$
  $RelevanceScore := (Loglike_T - Loglike_S)/Len$
  $InformativeScore := -AvgCNEntropy(\lambda_T, x)$
  $FinalScore := RelevanceScore * \theta + InformativeScore$
  if $(FinalScore > minScore)$ **then**
    $\mathcal{L}_x := QueryLabel(x)$
    $\mathcal{X}_T := \mathcal{X}_T \cup (x, \mathcal{L}_x)$
    $\mathcal{X}_M := \mathcal{X}_M \setminus x$
  end if
end for
$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
return λ_T

5. EXPERIMENT SETUP

5.1. Datasets

We conducted active learning experiments on both multi-accented and dissimilar accent cases. Multi-accented setup is based on GALE Arabic database. 1100 hours of Broadcast News (BN) is used as the source training data. It contains mostly Modern Standard Arabic (MSA) but also varying amounts of other dialects. LDC provided 30 hours of Levantine annotations on the Broadcast conversations (BC) portion of the GALE corpus. We assinged Levantine as our target accent and randomly selected 10 hours from LDC annotations and created our adaptation dataset. The remaining 20 hours of Levantine speech is mixed with 200 hours of BC data to create the Mixed dataset. This serves as our unlabeled dataset for active learning.

For dissimilar accent case, we chose English WallStreet Journal (WSJ1) as our source data. It contains 66 hours of read American English speech. We used British English version of the WSJ corpus (WSJCAM0) for adaptation. We randomly sampled 3 hours from WSJCAM0 for our adaptation

set. The remaining 12 hours of British English speech is mixed with 15 hours of American English from WSJ0 corpus to create our mixed dataset. The test sets, LM and dictionary are similar to our earlier setups in [15, 16]. Table 1 provides a summary of the datasets used.

Table 1. *Database Statistics.*

Dataset	Accent	#Hours	Ppl	%OOV
Arabic				
Training	Mostly MSA	1092.13	-	-
Adaptation	Levantine	10.2	-	-
Mixed	Mixed	221.9	-	-
Test-SRC	Non-Levantine	3.02	1011.57	4.5
Test-TGT	Levantine	3.08	1872.77	4.9
English				
Training	US	66.3	-	-
Adaptation	UK	3.0	-	-
Mixed	Mixed	27.0	-	-
Test-SRC	US	1.1	221.55	2.8
Test-TGT	UK	2.5	180.09	1.3

5.2. Baseline systems

We trained HMM-based speaker-independent systems on the training data using the Janus toolkit. They are Maximum Likelihood (ML) trained, context-dependent, fully-continuous systems with global LDA and Semi-Tied Covariance (STC) transform. More details on the front-end, training and decoding framework are explained in [17, 15]. We initially adapt our baselines systems on the relatively small, manually labeled, target adaptation dataset. We used semi-continuous polyphone decision tree adaptation (SPDTS) [15] for the supervised adaptation. The Word Error Rate (WER) of the baselines and supervised adaptation systems are given in Table 2.

Table 2. *Baseline and Supervised adaptation WERs.*

System	# Hours	Test WER (%)	
		SRC	TGT
Arabic			
Baseline	1100	46.3	53.7
Supervised Adapt	+10	51.4	52.1
English			
Baseline	66	13.4	30.5
Supervised Adapt	+3	21.0	17.9

6. IMPLEMENTATION DETAILS

We use the supervised adapted systems to select utterances from the mixed set for the goal of target accent adaptation. Our mixed sets were created by combining two datasets, American and British English or BC and Levantine Arabic. We evaluate 3 different data selection algorithms for our experiments: Random sampling, Uncertainty or informative sampling and relevance augmented uncertainty sampling. In each case, we select different amounts of audio data and mix it with the adaptation data. We then re-adapt the source ASR on the newly created dataset. For this second adaptation, we reuse the adapted polyphone decision tree from the supervised case, but we re-estimate the models on the new dataset using Maximum A Posteriori (MAP) adaptation.

In random sampling, we pick at random the required number of utterances from the mixed dataset. The performance of the re-trained ASR directly depends on the composition of source and target utterances in the selected subset. Thus, ASR re-trained on randomly sampled subsets will exhibit high variance in its performance. To avoid varying results, we can run random sampling multiple times and report the average performance. The other solution is to enforce that the randomly selected subset retains the same composition of source and target utterances in the mixed set. We use the latter approach for the results reported in this paper.

For uncertainty based sampling, we used average entropy calculated over the confusion networks (CN) as explained in section 2. We decode the entire mixed set and choose utterances that have the highest average CN entropy. In the case of relevance augmented uncertainty sampling, we use a weighted combination of relevance and uncertainty or informativeness scores for each utterance. The relevance score is derived from adapted and unadapted model cross-entropies with respect to the utterance. We calculate cross-entropy or average log-likelihood scores using the lattices produced during decoding. The uncertainty score is calculated using average CN entropy as before. We tuned the mixing weights on the English development set and we use the same weight (0.1) for all the experiments. We selected 5, 10, 15, 20 hour bins for English and 5, 10, 20, 40, 80 bins for Arabic. We choose utterances for each bin and combine it with the initial adaptation set, re-adapt the ASR and evaluate it on the target test set.

Table 3 shows WER of the oracle and select-all benchmarks for the two datasets. The oracle involves selecting all the target (relevant) data for human transcription, that we combined with source data to create the mixed dataset. The selected data is added to the initial adaptation set and used to re-adapt the source ASR. We note that in the case of Arabic, the source portion (BC) of the mixed dataset can have additional Levantine utterances, so oracle WER is not the lower bound for Arabic. Select-all involves selecting the whole mixed dataset for manual labeling. From Table 3, we can realize the importance of the relevance measure for active learning. In the case of Arabic, one-tenth of relevant data produces better performance on the target test set than the whole mixed dataset. The case is similar for English, where half of the relevant utterances help ASR achieve better performance than presenting all the available data for labeling.

Table 3. *Oracle and Select-all WERs.*

System	# Hours	Target WER
Arabic		
Oracle	$10 + 20$	48.7
Select-all	$10 + 221.9$	50.8
English		
Oracle	$3 + 12$	14.2
Select-all	$3 + 27$	14.9

7. ACTIVE LEARNING

The results for active learning for Arabic is shown in Figure 1. It is clear from the plot that the weighted combination of relevance and informative scores perform significantly better than uncertainty based score and random sampling techniques. We observe a 1.7% absolute WER reduction at the peak (40hours) for the weighted score when compared to the CN entropy based data selection technique. Also, with only 5 hours, the weighted score reaches WER of 49.5% while the CN-entropy based technique required 40 hours of data to reach a similar WER of 49.8%. Thus the combined score requires 87.5% less data to reach the same accuracy of CN-entropy based sampling. It is also interesting to note that our algorithm has identified additional Levantine data than the oracle from the generic BC portion of the mixed set which resulted in further WER reductions.

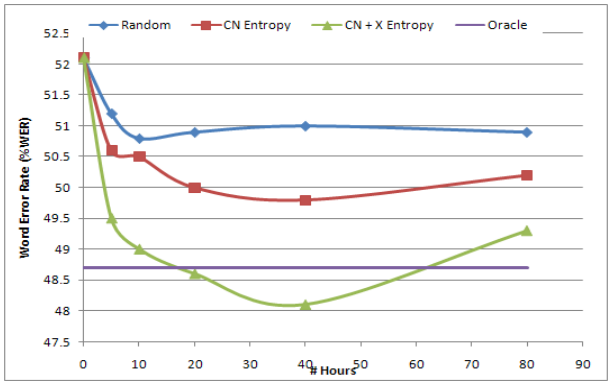


Fig. 1. *Active learning results for Arabic*

Figure 2 shows the equivalent plots for English. The combined score outperforms other techniques interms of the WER and reaches the performance of the oracle benchmark. It obtains similar performance with 10 hours of data (14.5%) compared to CN-entropy based technique at 20 hours (14.8%), thus achieving a 50% reduction in labeling costs.

8. ANALYSIS

In this section we analyze influence of relevance score in choosing the utterances that match the target data in both the setups. We plot the histogram of both CN-entropy and

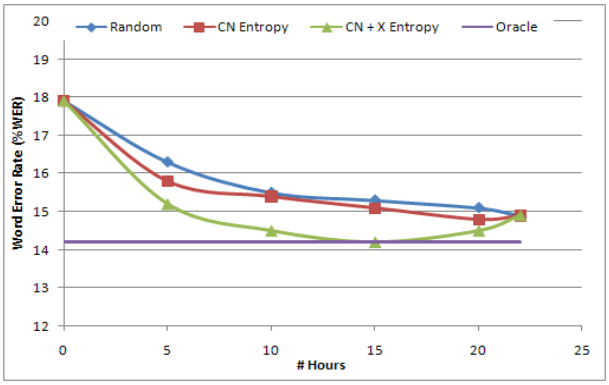
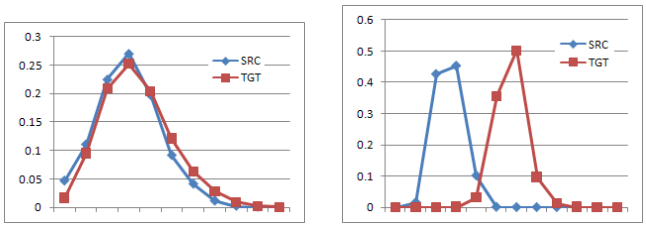


Fig. 2. *Active learning results for English*

weighted scores for each task. Figure 3 shows the normalized histograms for the American and British English utterances in the mixed set. We note that the bins for these graphs are in the ascending order of their scores. Data selection starts with the high-scoring utterances, hence the utterances from the right side of the plot are chosen first during active learning. Figure 3(a) shows the entropy scores for source (American English) and target (British English) are quite similar and the algorithm will find it harder to differentiate between relevant and irrelevant utterances based solely on uncertainty score. Figure 3(b) shows the influence of adding relevance scores to uncertainty scores. In this case, the target utterances have higher scores than source utterances and the algorithm chooses relevant ones for re-training the ASR.

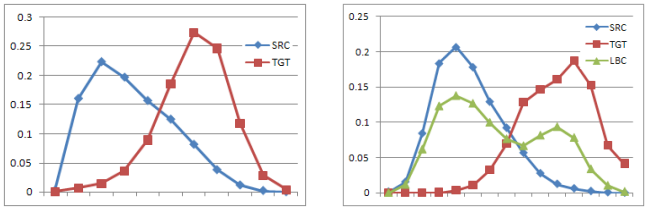


(a) Entropy (b) Weighted Score

Fig. 3. Histogram of source and target scores for English.

Figure 4 shows similar plots for Arabic. The distinction between CN-entropy and the weighted score in source/target discrimination is less clear here compared to English plots. However, we can still see that target utterances achieve better scores with weighted combination than the CN-entropy score. We observed many of the utterances from 'LBC_NAHAR' shows, part of the BC portion of the mixed set, ranked higher in the weighted score. The plot of LBC scores in the histogram shows these utterances from the BC portion have high scores in the weighted case. They are recording of the 'Naharkum Saiid' (news) programmes from Lebanese Broadcasting Corporation originating from the Levantine region and likely to have Levantine speech. This observation shows

that the relevance score identifies additional Levantine speech from the BC utterances.



(a) Entropy (b) Weighted Score

Fig. 4. Histogram of source and target scores for Arabic.

9. CONCLUSION AND FUTURE WORK

We introduced cross-entropy based relevance score to augment uncertainty sampling based active learning in speech recognition for the goal of accent adaptation. We showed that our algorithm can achieve similar accuracy with only 12.5-50% of the utterances selected by uncertainty based technique for the two tasks in Arabic and English accents. We analyzed influence of relevance score in data selection and showed that its capable of identifying appropriate utterances that match the target set, from the untranscribed data with varying accents. Future work includes incorporating semi-supvervised training [16] incombination with active learning for further improvements. We also plan to explore the use of relevance scores in unsupervised discriminative training [18].

10. REFERENCES

[1] Burr Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[2] Katrin Tomanek and Fredrik Olsson, "A web survey on the use of active learning to support annotation of text data," in *Workshop on Active Learning for NLP*, Stroudsburg, PA, USA, 2009, HLT '09, pp. 45–48.

[3] Gökhan Tür, Dilek Z. Hakkani-Tür, and Robert E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[4] Giuseppe Riccardi and Dilek 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.

[5] Kai Yu, Mark J. F. Gales, Lan Wang, and Philip C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[6] 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 & Language*, vol. 24, no. 3, 2010.

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

[8] Dilek Z. Hakkani-Tür, Giuseppe Riccardi, and Allen L. Gorin, "Active learning for automatic speech recognition," in *ICASSP*, 2002, pp. 3904–3907.

[9] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou, "Active learning by querying informative and representative examples," in *NIPS*, 2010, pp. 892–900.

[10] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, 2000.

[11] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh, "Sample selection bias correction theory," in *ALT*, 2008, pp. 38–53.

[12] John Blitzer and Hal Daumé III, "ICML tutorial on domain adaptation," `http://adaptationtutorial.blitzer.com`, June 2010.

[13] Steffen Bickel, Michael Brückner, and Tobias Scheffer, "Discriminative learning under covariate shift," *Journal of Machine Learning Research*, vol. 10, 2009.

[14] Robert C. Moore and William Lewis, "Intelligent selection of language model training data," in *ACL (Short Papers)*, 2010, pp. 220–224.

[15] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz, "Enhanced polyphone decision tree adaptation for accented speech recognition," in *INTERSPEECH*, 2012.

[16] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz, "Enhanced polyphone decision tree adaptation for accented speech recognition," in *MLSLP Symposium*, 2012.

[17] Florian Metze, Roger Hsiao, Qin Jin, Udhyakumar Nallasamy, and Tanja Schultz, "The 2010 cmu gale speech-to-text system," in *INTERSPEECH*, 2010.

[18] Xiaodong Cui, Jing Huang, and Jen-Tzung Chien, "Multi-view and multi-objective semi-supervised learning for hmm-based automatic speech recognition," *IEEE Transactions on Audio, Speech and Language processing*, vol. 20, no. 7, pp. 1923–1935, 2012.