SUPERVISED AND UNSUPERVISED ACTIVE LEARNING FOR AUTOMATIC SPEECH RECOGNITION OF LOW-RESOURCE LANGUAGES

Ali Raza Syed[⋆] *Andrew Rosenberg*[⋆] *Ellen Kislal*[†]

[⋆] The Graduate Center, CUNY, New York, NY, USA
[†]IBM TJ Watson Research Center, Yorktown Heights, NY, USA

ABSTRACT

Automatic speech recognition (ASR) systems rely on large quantities of transcribed acoustic data. The collection of audio data is relatively cheap, whereas the transcription of that data is relatively expensive. Thus there is an interest in the ASR community in active learning, in which only a small subset of highly representative data chosen from a large pool of untranscribed audio need be transcribed in order to approach the performance of the system trained with much larger amounts of transcribed audio. In this paper, we compare two basic approaches to active learning: a supervised approach in which we build a speech recognition system from a small amount of seed data in order to make the selection of a limited amount of additional audio for transcription, and an unsupervised approach in which no intermediate system recognition system built from seed data is necessary. Our best unsupervised approach performs quite close to our supervised approach, with both outperforming a random selection scheme.

Index Terms— supervised active learning, unsupervised active learning, limited-resource automatic speech recognition, active learning

1. INTRODUCTION

Active learning is a special case of machine learning in which the learning algorithm is allowed to query an information source for additional ground-truth annotations during the learning process. In the case of speech recognition, the training algorithm identifies segments of untranscribed audio for which a transcript is requested. We are working in the context of the IARPA Babel challenge in which one hour of transcribed seed data is provided in a language and participants are allowed to request transcripts for an additional two hours of transcriptions. The resulting three hours of transcribed data are then used to build an ASR system.

A strong unsupervised approach to identify "good" segments for annotation is valuable for a number of reasons. First, when developing a seed set of annotated data, it would be helpful to prioritize representative or informative data points (for ASR, utterances). This would lead to improved seed models and, thereby, superior active learning selection criteria. Second, unsupervised active learning has the potential to save computational resources. Training ASR models is a resource-intensive task. To train an end-to-end ASR system can take days. Because of this, seed data is frequently used to build a weaker ASR system to generate selection criteria, with a strong, state-of-the-art (i.e. "all the bells and whistles") system built only on the larger, augmented training set. Unsupervised data selection avoid the resource requirements of the first-pass training altogether. Third, supervised active learning has the weakness of not necessarily being able to select samples that represent classes unseen in the seed data; unsupervised active learning is not hindered by the same limitation.

In this paper we examine unsupervised approaches to active learning, in which no model is required to be built from the original seed data in order to decide which of the remaining segments of speech would be most usefully transcribed. We contrast the unsupervised approach with a more traditional supervised scheme in which we build an ASR system from the seed data and use that model to decode the remaining segments as the first step in making the selection of the remaining data for which to request a transcription. The work is evaluated under the IARPA Babel evaluation framework for Swahili using the one hour seed data (transcribed data) and selection pool (untranscribed data) as described by IARPA-babel202b-v1.0d ALP ("Active Learning Pack") distribution. We show that we can come quite close to the supervised approach with a well-designed unsupervised one; the supervised WER on a development set was 67.8% whereas the best unsupervised WER on the same set was 68.0%. Both of these outperformed a random selection which had a WER of 69.6%.

In this paper, we first describe our baseline system which is a supervised active learning Swahili speech recognition system built with one hour of transcribed seed data plus two additional hours of transcribed acoustic data selected by maximizing grapheme entropy per segment (Section 3.1). We then describe various approaches to unsupervised active learning using the same one hour seed data to select an additional two hours of data for transcription (Section 3.2). We then present results (Section 4) and conclude (Section 5).

2. RELATED WORK

Traditional active learning schemes for ASR systems employ uncertainty sampling techniques [1]. This involves an iterative process: an ASR is trained on transcribed utterances and used to label untranscribed utterances; a confidence score is calculated for each labeled instance; the utterance with lowest confidence is selected for transcription then added to the ASR's training set before re-training the ASR and repeating the process. Thus each iteration tries to obtain only examples of utterances that the ASR has mis-labeled. This framework requires a measure of uncertainty or confidence to be assigned to the examples. In [2], Tur et al. developed a method for assigning confidence scores to untranscribed utterances based on the ASR lattice output. In [3], rather than using a threshold confidence score to select utterances, Riccardi et al. filter the confidence scores through an informativeness function to rank the utterances before selecting a subset of utterances. They found that a function which penalizes both high and low scores works better, thus avoiding utterances where the ASR is either highly confident or highly unconfident. In [4], Yu and Gales focus on improving ASR performance by targeting discriminative training where performance is sensitive to transcription quality. The authors develop a scheme for directed manual transcription of a small portion of poorly recognized data which is shown to improve performance relative to automatic transcription techniques. The work by Fraga-Silva et al. [5] is close to our domain as their experiments are conducted with a low-resource language from the IARPA-Babel corpus using the same settings (Babel Active Learning task). The authors investigate a number of metrics for selection and find that HMM-state entropy and vocabulary size correlate best with WER. They use HMM-states to model the acoustic space and yield selections with highest entropy with a greedy algorithm.

More recent work has focused on the use of submodular functions [6] for data selection. A submodular function F is a value function over sets satisfying the following property for any $S \subseteq S'$:

$$F(S \cup \{u\}) - F(S) \geq F(S' \cup \{u\}) - F(S'). \quad (1)$$

Informally, submodular functions capture the notion of "diminishing returns": the marginal benefit of adding an example u to a set S is at least the marginal benefit of adding the same example to a larger superset $S' \supseteq S$. The property of submodularity guarantees that, for a monotonically non-decreasing F, a greedy algorithm with a cardinality constraint will yield a near-optimal selection. Formulating data selection as optimization of a submodular objective function is particularly well suited for active learning tasks where data must be selected in batches constrained by a budget. A number of submodular objective functions have been considered and proposed in the literature. In [7], the facility location function, which measures similarity of a subset S to the entire pool P, was used to select a representative subset. In contrast to the graph based facility location function, Wei et al. [8] introduce a multilayer feature based submodular function. This objective seeks a representative selection over one feature space while also attaining a diverse selection by considering interactions between a high- and a low-level feature space. In [9], Chen et al. also formulate a submodular objective to attain both a representative and diverse selection by considering a (single-layer) feature based function over the acoustic characteristics.

3. ACTIVE LEARNING

Following the IARPA Babel evaluation scenario, all active learning selection strategies use the same, predefined, one-hour seed set of data, and select an additional two hours. As a baseline, we construct a random selection of two-hours of data. In this approach, the selection pool is segmented using a VAD system. Based on this segmentation, segments are selected at random, with no additional selection criteria. This random selection of segments led to a WER of 69.6%. The rest of this section describes the supervised and unsupervised selection criteria explored.

3.1. Supervised Active Learning

The first step in selecting a two-hour subset of data from the 30-hour pool of candidates is to build a multilingual-feature context-dependent Gaussian mixture model (GMM) from the one hour of Swahili seed data using a graphemic lexicon. We used the IBM Attila toolkit [10] for building this model.

The multilingual feature vector was a 62-dimensional bottleneck features from a DNN trained on the 11 Babel base period and option period 1 language data appended to the standard 40-dimensional IBM PLP+LDA+STC speaker-independent feature vector. Additional details of this feature generation approach can be found in [11] .

In order to make the segment selection, we do a consensus decode [12] of the untranscribed segment pool and calculate a grapheme-based entropy for each segment. The grapheme-based entropy is approximated by splitting the words appearing in the consensus network into their constituent graphemic strings and incrementing the grapheme count in a histogram by the probability weighting for that word. The grapheme pdf is normalized by the total counts and the entropy of the pdf is calculated.

We found that ignoring speaker identity and greedily selecting segments based on their entropy out-performed a scheme in which segments were selected in a "round-robin" fashion by speaker (the segment having the highest entropy for a given speaker was chosen as we rotated through speakers.) Segments were required to have a duration greater than 0.75 seconds and not to come from the beginning 100 seconds or after the 500th second of a speaker turn in order to be considered in the selection. The supervised active learning approach led to a WER of 67.8% on the development Swahili test set.

3.2. Unsupervised Active Learning

In unsupervised active learning, we investigated methods of selecting a high quality 2 hour subset of the ALP pool without the use of transcripts or the output of a first-pass ASR system. Avoiding the ASR training and decoding process allowed for faster turnaround time in experimentation. Also, prior work in the active learning literature indicates that bootstrap (first-pass) ASR systems typically require far in excess of 1 hour of training data. We are skeptical that the confidence scores from a system built on such a small amount of data will be informative.

Our selections are based on two methods: 1) selecting based on speech rate of utterances, 2) selecting representative samples based on acoustic features.

Selection based on speech rate. This method uses the speaking rate of an utterance as a proxy for the phone or word density of that utterance. The assumption here is that a selection that contains more phones, and thus, more words, would provide more training instances for both acoustic model training and more tokens for language modeling. One risk is that this approach introduces a bias toward rapid speech.

Selecting high density utterances yields a maximum number of tokens for ASR training. Speech rate estimation was performed using signal processing techniques based on syllable nuclei detection. We explored two publicly available speech rate estimators: AuToBI [13, 14] and an implementation of *mrate* [15]. We found that the AuToBI implementation of the Villing 2004 algorithm was a more effective measure of speaking rate on the one-hour training pool, and, thus, used this approach for selection.

Selection of representative samples. Here, we employ methods for selecting utterances which are most representative of the acoustic feature distribution of the ALP pool. This set of approaches is based on the assumption that a maximally representative selection of the overall acoustic feature space represents an optimal set of ASR training data. This should have only minimal impact on language modeling, though we expect the resultant "graphone" distribution to be more or less representative if the acoustic feature space is effectively sampled. We evaluate two selection schemes: distribution matching and subset selection using a facility location function. This, and subsequent, acoustic feature analyses are performed on the multi-lingual features described in Section 3.1.

To facilitate analysis, we discretize the acoustic feature space by learning a k-means codebook $C_1, C_2, ..., C_k$ over the utterances with an encoding function g. For an utterance u with m frames, we find the centroid closest to each frame and assign a vector $g(u) = \frac{1}{m}[c_1(u) \cdots c_k(u)]$, where $c_i(u)$ counts the number of occurences of centroid C_i. For each of the selection functions, we generate a codebook with $k = 256$ entries.

KL-divergence Minimization: This distribution matching approach is a sampling scheme for selecting utterances such that the feature distribution of selected utterances (sample distribution) matches the feature distribution of the ALP selection pool (reference distribution). We use KL-divergence between the sample distribution and the reference distribution as an the objective function. We explore two algorithms to perform distribution matching. The first algorithm employed a greedy method to grow the selection iteratively such that the updated selection had the minimal KL-divergence to the reference at the end of each iteration. This algorithm is fairly slow, as the KL-divergence between the sample distribution and reference-distribution must be recalculated each time an utterance is added to the sample distribution. The second algorithm employed a Knapsack-problem formulation. In this, each utterance is assigned a value inversely proportional to its KL-divergence from the reference. Then, the goal was to fill a 2 hour knapsack with the maximum total value of utterances, which was optimized using the 0-1 knapsack dynamic programming formulation. While this is much faster, it has a tendency to oversample common features and undersample less frequent areas of the acoustic feature space.

Subset selection: The facility location function (Equation 2) is a submodular function which measures the similarity of a selection S to the remainder of the ALP pool U [7].

$$f(S) = \sum_{u \in U} \max_{s \in S} w_{u,s} \tag{2}$$

Then we build a graph with edge weights $w(u, s)$ measuring the similarity between $g(u)$ and $g(s)$, where $g(.)$ is the encoding function mentioned above. Our experiments used cosine similarity for the weight function $w(u, s)$. Optimizing over the function (2) using a greedy method then yields a representative subset.

As in the supervised selection (cf. Section 3.1), we had initially explored methods whereby we select the best segment within for each speaker and rotated across speakers in a round robin fashion. We found, however, in the unsupervised context as well, that omitting this speaker-balancing criterion led to improved selections.

Selection with diversity reward: One limitation of the facility location method, similar to the knapsack formulation of the KL-divergenge approach, is that it can oversample the most common features (i.e. central data points) while omitting less frequently observed areas of the feature space. Thus, we include a criteria to promote a more diverse selection set. We investigate extending the facility location method to yield a selection which was representative of the pool while selecting a diverse set of utterances. This was performed by regularizing the facility location objective with a reward for diversity based on cluster membership of the utterances (cf. Equation 3). Since the cluster diversity function is also submodular [16], this objective is optimized using the greedy method.

$$f(S) = \sum_{u \in U} \max_{s \in S} w_{u,s} + \lambda \sum_{i} \delta(S \cap C_i \neq \emptyset) \tag{3}$$

The second term in this equation counts the number of codebook entries that are represented in the utterance. Thus, utterances that contain a large range of acoustic features are preferred to those which contain fewer.

4. RESULTS AND DISCUSSION

We evaluate the WER on the Swahili development set. This is a set of 10 hours of speech. Results across all selection methods are presented in Table 1.

Method	WER
Knapsack KL-div	73.0
Facility Location	69.8
Random Selection	69.6
Speech rate	69.1
Greedy KL-div	68.5
Facility Location with Cluster diversity	68.0
Grapheme Entropy	67.8

Table 1. *Swahili WER from 3-hr Active Learning ASR*

We find that the supervised selection approach to result in the most effective active learning selection, yielding a WER of 67.8. However, this is only 0.2% better than the best unsupervised approach, Facility Location with Cluster diversity.

We also notice how poor a selection criteria the knapsack-based formulation of the KL-divergence minimization approach is. This results in a training set that performs quite a bit worse, 3.4% absolute, than Random selection. While the Facility Location approach performs better, it also performs worse than Random by 0.2%. Both of these techniques take an approach by which the "centrality" of each utterance is measured with the selection set being constructed from only the most typical utterances. Neither includes any measure of diversity. This suggests that when constructing a 3-hour training set, diversity is an important criteria. When we turn our attention to the two corresponding approaches, Greedy KL-div, and Facility Location with Cluster Diversity, we find much better performance, 68.5, and 68.0 respectively. Recall that Greedy KL-div iteratively constructs a selection set whose acoustic feature distribution matches the overall selection set. By updating the KL-divergence between the selection and target, this approach has the ability to select utterances that, when viewed in isolation, are outliers with respect to the overall feature distribution, but when viewed with respect to the current selection set are representing an under-sampled region of the feature space. The Cluster Diversity measure augments the facility location approach by encouraging the selection of utterances that contain more diverse acoustic features.

The speech rate selection criteria outperforms the random selection but is not competitive with the diversity sensitive distribution matching approaches.

The distribution matching approaches are based on the assumption that ASR training data should have a representative distribution of the acoustics of the language being recognized. This may be an effective approach for constructing an effective acoustic model, mapping from the acoustic feature space to the target phone (or here, graphone space).

However, ASR training involves not only conversion from acoustics to the phonological space, but also effective pronunciation and language modeling. Thus, there is a slight difference between optimizing the representativeness of the acoustic feature space and ASR performance. So while submodular functions provide an effective way to optimize the representativeness of the selection set with respect to the acoustics, there is a mismatch between this optimization and optimizing WER. Likely through a greater match between the selection criteria and WER objective, the supervised, grapheme entropy approach represents a more effective active learning criteria.

5. CONCLUSION

This work is performed as part of the IARPA Babel program to develop speech recognition and keyword search for low-resource languages. We present results on a variety of approaches to active learning for automatic speech recognition. While we find that we can identify a training set that outperforms random selection by 1.8% WER, it is difficult, with three hours of training data, to make a very large improvement.

We find that supervised active learning, where a first-pass ASR system is trained and the hypotheses of the system are used in selecting a larger set of annotated data for training to be effective. However, we also demonstrate the use of unsupervised selection approaches which do not require a first-pass ASR system; the best performing of which is only 0.2% worse than a supervised approach. The most effective unsupervised approaches require that the selection set be both informative, but also diverse. Future work will investigate the impact of combining effective supervised and unsupervised selection approaches and investigating supervised criteria involving multiple ASR training passes.

6. ACKNOWLEDGEMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. REFERENCES

[1] Burr Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, pp. 1–114, 2012.

[2] Dilek Hakkani-Tur, Giuseppe Riccardi, and Allen Gorin, "Active learning for automatic speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing*. 2002, pp. IV–3904–IV–3907, IEEE.

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

[4] Kai Yu, Mark 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.

[5] Thiago Fraga-Silva, Jean-Luc Gauvain, Lori Lamel, Antoine Laurent, Viet-bac Le, and Abdel Messaoudi, "Active Learning based data selection for limited resource STT and KWS," in *INTERSPEECH*, 2015, pp. 3159–3163.

[6] G. L. Nemhauser, L. a. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions-I," *Mathematical Programming*, vol. 14, pp. 265–294, 1978.

[7] Hui Lin and Jeff Bilmes, "How to Select a Good Training-data Subset for Transcription : Submodular Active Selection for Sequences," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 18–21.

[8] Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes, "Unsupervised submodular subset selection for speech data," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2014, pp. 4107–4111, IEEE.

[9] Nancy F Chen, Chongjia Ni, I-fan Chen, Sunil Sivadas, Van Tung Pham, Haihua Xu, Xiong Xiao, Tze Siong Lau, Su Jun Leow, Boon Pang Lim, Cheung-chi Leung, Lei Wang, Chin-hui Lee, Alvina Goh, Eng Siong Chng, Bin Ma, and Haizhou Li, "Low-resource keyword search strategies for Tamil," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[10] H. Soltau, G. Saon, , and B. Kingsbury, "The ibm attila speech recognition toolkit," in *IEEE Workshop on Spoken Language Technology*, 2010.

[11] Z. Tuske, P. Golik, D. Nolden, R. Schluter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages," in *INTERSPEECH*, 2014, pp. 1420—1424.

[12] 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, pp. 373–400, 2000.

[13] Andrew Rosenberg, "Autobi-a tool for automatic tobi annotation.," in *INTERSPEECH*, 2010, pp. 146–149.

[14] R. Villing, J. Timoney, T. Ward, and John Costello, "Automatic blind syllable segmentation for continuous speech," *IET Conference Proceedings*, pp. 41–46(5), 2004.

[15] D. Wang and S.S. Narayanan, "Robust speech rate estimation for spontaneous speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2190–2201, 2007.

[16] Adarsh Prasad, Stefanie Jegelka, and Dhruv Batra, "Submodular Maximization and Diversity in Structured Output Spaces," in *NIPS*, 2014.