2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 149 - 153

Efficient Use of Training Data for Sinhala Speech Recognition using Active Learning

Thilini Nadungodage[*1], Ruvan Weerasinghe[*2], Mahesan Niranjan[†3]
[*]*Language Technology Research Laboratory*
University of Colombo School of Computing, Colombo 00700, Sri Lanka
[†]*School of Electronics and Computer Science*
University of Southampton, Highfield, Southampton SO17 1BJ, UK

[1]hnd@ucsc.lk [2]arw@ucsc.lk [3]M.Niranjan@Southampton.ac.uk

Abstract*— Automatic Speech Recognition is an area which requires a large amount of training data. Collecting such quantities of data involves significant time and cost owing to the tedious nature of collecting speech recordings and manual nature of transcribing it. For a low resourced language such as Sinhala, collecting a sufficient data set is a major problem. To address this issue we used the Active Learning technique from the Machine Learning paradigm which is applied to many tasks such as information retrieval. Our experiment using a simple Sinhala speech corpus shows that through the use of Active Learning, the amount of utterances that need to be transcribed can be reduced by some 42% to achieve the same accuracy as using the whole data set without such a strategy. This suggests that Active Learning techniques can be successfully applied to make optimal use of scarce resources for speech recognition for new languages.

***Keywords*— Automatic Speech Recognition, ASR, Information extraction, Low Resourced Languages, Active Learning, Word posterior probabilities, Confidence scoring, Natural Language Processing, NLP, Sinhala**

I. INTRODUCTION (HEADING 1)

Automatic speech recognition (ASR) has attracted much interest in the research community over several decades. The attraction of building automatic speech recognition devices to substitute human operators, dictation systems that can replace keyboards and voice control of machines have resulted in significant industrial research and development in the area. In limited applications, considerable advances have been made and dialogue systems that can interact with human inquiries over telephone have now been deployed. Speech being a natural medium of human communication also encourages us to seek to understand how linguistic and paralinguistic information is encoded in the signal, and how human perception is able to normalize unwanted information and extract the underlying message. While rule based techniques dominated speech recognition research during its very early days of development, statistical techniques that capture uncertainties in the speech at the signal processing, acoustic modelling and linguistic stages via probabilistic modelling techniques offer state-of-the-art performance nowadays.

Much of the above advance, however, is restricted to languages for which large corpora are available. English and other European languages have seen very rapid advances in speech research due to availability of funding and advanced research environments in the USA and the EU. More recently, Chinese and Arabic have seen rapid increase in interest, and resources in the form of speech corpora are available for these languages. Languages spoken by relatively smaller populations, and primarily in developing countries, have not seen the same level of interest and consequently are data-poor. Sinhala is one such example.

This paper is about the use of Active Learning, a technique that allows to gather training data incrementally in a way that maximizes the information content of the resulting dataset. It is based on the premise that, in a resource-poor environment, seeking to increase the amount of resources at random can be wasteful. Active learning starts from a small dataset, trains a model and seeks to acquire new data based on how informative would that data be, with respect to what has been learned so far. Thus scarce resources can be deployed in an optimal fashion.

The remainder of this paper is as follows. Section II, A description of the Active learning approaches and related works. Section III, description of the approach we used and Section IV, algorithms used in the Active learning computations. Section V, the experiment and the results. And finally Section VI gives the conclusions of the experiment.

II. RELATED WORK

Data sparseness has been a major problem in the machine learning field from the beginning. To answer this problem, Active Learning (AL) was introduced to the context of machine learning. The work presented in [1] can be considered as one of the first attempts in using AL in machine learning and it has been an inspiration for many subsequent AL applications in various sets of machine learning fields [7]. Also it has been used in number of Natural Language processing tasks such as information extraction, named entity recognition, text categorization, part of speech tagging, and parsing among many others [11]. Active learning techniques have also been used in acoustic modelling for spoken language understanding in the recent years [4]. Details on the theory behind Active Learning and its application in various machine learning problems can be found in [14] and [15]. In [13], the author present how

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

Active Learning can be used in speech recognition applications. There are many papers in the literature that illustrates various applications of AL techniques in the speech recognition domain [9], [16], [19], [21].

In speech recognition applications that use AL techniques, Confidence Measures plays a major roll. These measure how confident a model is of its recognition output for a particular utterance. This helps to determine the novel and more informative data sample from a given data set. We can say that, the more uncertain (less confident) a model is about its output, the more that utterance contains novel information.

According to [2] and [8] Confidence Measuring techniques can be mainly divided in to two categories. One category builds confidence predictor features based on acoustic and language model information collected during decoding. Some of these features are: Normalized Likelihood Score, N-best related features, Word graph related features, Acoustic Stability, Hypothesis Density, Language Model related features, Parsing related, Duration related, and Log Likelihood Ratio related features. The second category uses posterior probability based confidence measures computed either during decoding or in a post-processing step on N-best lists or word graphs. From these many techniques that can be used in confidence measurement, [3] and [18] shows that word posterior probabilities outperform the other confidence measurement techniques.

Experiments have shown that using Active learning techniques to extract the most informative data has helped to reduce the effort and resources needed to transcribe speech data to achieve the same performance from a speech recognition model built with a whole data set. A simple experiment described in [5] shows that the same accuracy as using the whole data set can be achieved by using 27% less data by employing AL to select the most informative data for training. Similarly [13] shows that by using Active Learning, the amount of labelled data needed for a given word accuracy can be reduced by more than 60%. The authors of [12] show that by combining Active Learning techniques and Unsupervised Learning, the amount of labelled data needed for a given word accuracy can be reduced by 75%. [6] demonstrate that for a speech corpus of Spontaneous Japanese, they only required 63 hours of data to achieve a word accuracy of 74% using Active Learning, where standard training required 97 hours of data.

III. Approach

Our goal was to extract the most informative speech data from an existing speech database so that we can reduce the resources spent for the transcription process. To reduce the transcription effort we used the uncertainty-based active learning methods as described in [14]. We selected the data which the recognizer is most uncertain for transcription and left out the data which the recognizer is more certain about.

For this an initial speech recognizer has to be developed. We randomly selected a small sample of acoustic data and transcribed it to train the initial speech recognition model. Using this initial recognizer we recognized the remaining data which are the candidates for the transcription process. Then we computed the confidence scores of the recognized utterances to determine which utterances the recognizer is most certain about and which the recognizer is uncertain about. We selected the most uncertain utterances for manual transcription. We repeated this process until the most informative data set was extracted. The whole process can be described using the following steps:

1. Transcribe an initial data set (S_I)
2. Using S_I , train an Acoustic model (AM_i) and a Language model (LM_i), where i is the iteration number
3. Using AM_i and LM_i recognize the remaining candidate utterances (S_R)
4. Compute the word confidence scores and then utterance confidence scores for each utterance in SR
5. Select the K utterances with the smallest confidence scores from S_R
6. Manually transcribe the selected K utterances (S_i)
7. Add S_i to S_I. So,
$$S_I = S_I + S_i$$
$$S_R = S_R - S_i$$
8. Stop if the expected word accuracy is achieved, else repeat from step 2

IV. Confidence Score Computation

There are many confidence measurement computation methods presented in the literature. As mentioned in section II, experiments have shown that using word posteriors as the confidence measurement has proven to be more effective than other methods. So, in our experiment we used the word posterior probabilities as our confidence score. We used the Simple Accumulation method presented in [2] to improve the performance of the confidence scores.

A. Word Posterior Calculation

As described in [17], we used N-best lattices to compute the word posterior probabilities. We got N number of best recognition hypotheses for each candidate utterance and used the word graph generated by these hypotheses to compute the posteriors. The posterior probability of each word depends on the acoustic and language model probabilities of that word in every hypothesis.

B. Simple Accumulation Method

When computing word posterior probabilities, the acoustic and language model probabilities are obtained from equal words that start and end at the same time. However, there can be equal words in different hypotheses which slightly differ in their starting and ending times. So, in order to get a better confidence score, the posterior probabilities of equal words with overlapping time intervals were accumulated and the new values were considered as the confidence score of that particular word.

C. Utterance Confidence Score

There are several different approaches to obtain the utterance confidence scores using the word confidence scores [3]. One way is to get the arithmetic mean of the confidence scores of the words of that hypothesis as the utterance

confidence score. Another approach is to get the confidence score of an utterance as the product of the confidence scores of the words that it contains. Utterance confidence can also be computed using functions such as geometric-mean or min. Here we have used the arithmetic mean of the word posterior probabilities as the utterance confidence score.

V. Experiments and Results

We have done an experiment to show that Active Learning techniques can be used for Sinhala speech to reduce the amount of data that should be transcribed to get the accuracy that can be achieved from a whole dataset. We used a simple and small Sinhala continuous speech corpus that has been used to build a baseline Sinhala ASR system [10].

The full corpus contained 3,053 utterances (31,625 words) from a single speaker. We used the Hidden Markov Model Tool Kit (HTK) developed by the Cambridge University, UK [20] to build the acoustic models and the language models.

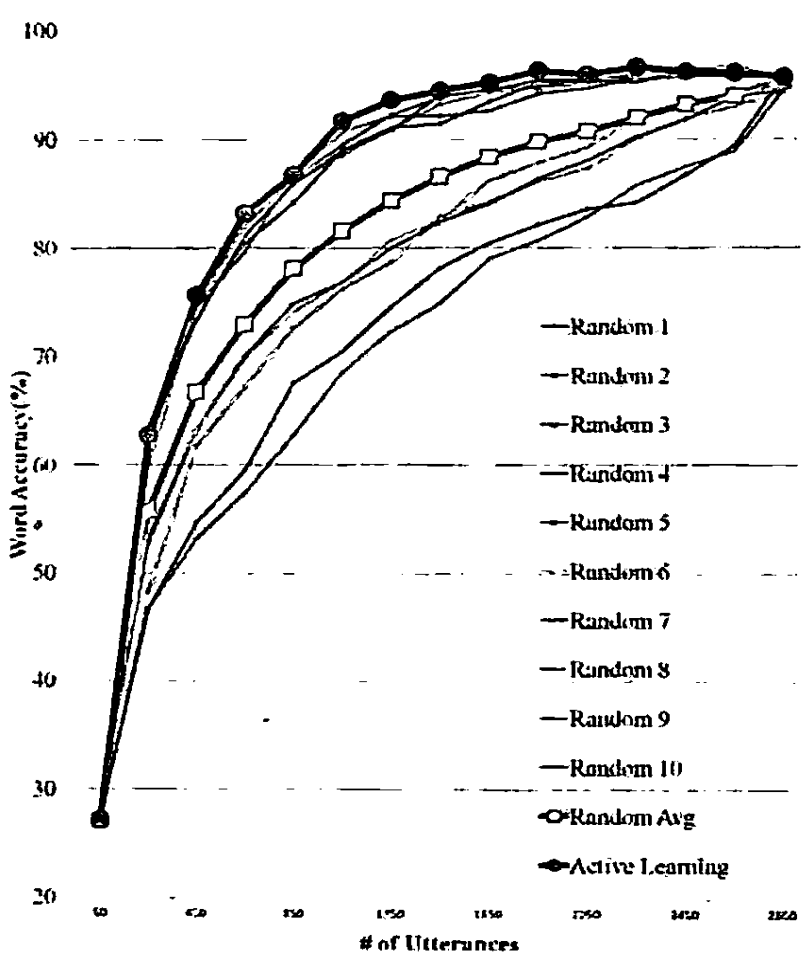

Figure 1: Increase in recognition accuracy as a function of number of training utterances. Active learning (black line) has a faster rise than random inclusion of data (grey line) because it selects novel data for inclusion into the training set. The graph for random increase in training set is averaged over 10 runs. Performances of individual random runs are also shown (light grey lines).

A. Training and Testing Data

We selected 2,947 utterances as the training data (30,796 words) and set aside 106 utterances (829 words) as test data. The initial set of transcribed utterances, which is used to train the initial acoustic and language models, consists of 50 utterances (403 words). The additional set of transcription candidate utterances consists of 2,897 utterances (30,393 words).

B. Experiment

Our experiment consisted of two parts, one using the Active learning data selection process and the other one using a Random data selection process, to measure how Active learning performs compared to random data selections. We selected 200 utterances in each step for transcription for both active learning and random setups.

1) *Active Learning Data Selection:* For the Active Learning process we started with the initially transcribed data set and built the acoustic and language models. Then we input the remaining candidate utterances to the trained model and got the 10 best output hypotheses for each utterance. We then computed the word confidence scores for each word in the created word graph and calculated the confidence scores of the best hypothesis for each utterance. We then selected the 200 utterances with the lowest confidence scores for transcription for training in the next step. We combined this selected data set with the initial data set and rebuilt the models using this combined data set. We measured the accuracy of the test set in each step and repeated the process until the accuracy from the whole data set was achieved.

2) *Random Data Selection:* For the Random data selection process also we started with the initially transcribed data set and trained the acoustic and language models. Next we randomly selected 200 utterances from the remaining candidate utterances and transcribed them and added them to the training set in the next step. We measured the accuracy of the test set in each step and repeated the process until the accuracy from the whole data set was achieved.

Since we couldn't prove the behaviour of the data from only one random data selection process, we performed the random data selection process 10 times and calculated the average accuracy values from each step in order to get a more representative value.

# of Utterances	Vocabulary Size	
	Active Learning	Random (Averaged)
50	262	262
250	1466	1003
450	2195	1531
650	2746	1945
850	3072	2293
1050	3395	2602
1250	3617	2862
1450	3809	3069
1650	3925	3270
1850	3998	3340
2050	4030	3594
2250	4048	3735
2450	4061	3816
2650	4067	3938
2850	4067	4036

Table 1: Vocabulary Sizes of the Active Learning and Random Data Selection Processes

3) Results: We obtained the accuracy for the test data set using each model built at every step. Figure 1 show how the accuracy increases in Active learning and Random data selection processes. The best performance with Random data selection was achieved using almost all the training data (2850 utterances). We achieved the same word accuracy (95.17%) with Active learning data selection using only 1650 utterances. This is 42% less data than the Random process. This shows that Active learning data sampling is more effective in the ex- traction of important information than random data sampling.

We further analyzed the vocabulary distribution during the training processes. Table I presents the vocabulary sizes of each step during the process. It shows that the Active learning process has increased the vocabulary size faster and in big quantities than the averaged Random learning process.

Figure 2 presents how many new (previously unseen) words are added to the vocabulary during each step of the process. It clearly shows that the Active learning process pulls new words immediately so the vocabulary grows faster making it learn new phenomena faster.

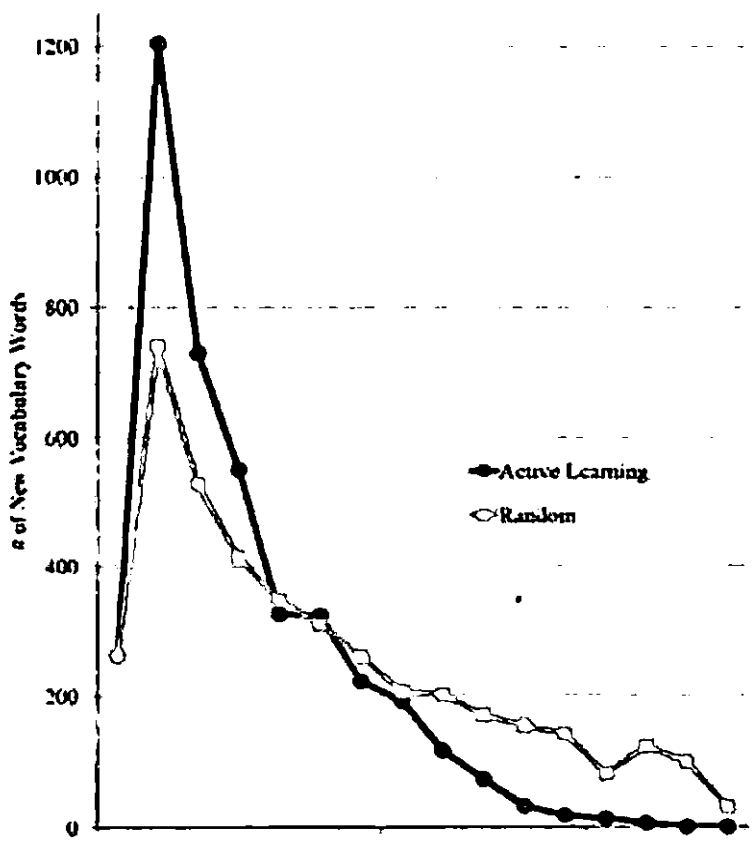

Figure 2: Behaviour of # of novel vocabulary words as a function of number of training utterances. Active Learning (black line) extracts novel words faster than the average Random (grey line) process.

VI. Conclusions

We have shown that active learning can be used for efficiently increasing the size of the training data set of a Sinhala speech recognition system. Low confidence recognition by a trained recognizer is used to select new data for inclusion in the training set, and subsequent re-training of the recognizer. For resource-poor languages such as Sinhala, this is a highly desirable strategy for quickly designing deployable speech recognition systems. It helps developers of speech recognizers, especially for these new languages, circumvent the large investment necessary for constructing speech corpora in the traditional (random selection) way.

VII. Acknowledgements

The authors would like to acknowledge the National Research Council (NRC) of Sri Lanka for funding this research. They are also indebted to the research team members of the Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka, for assisting in numerous ways.

References

[1] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[2] T. Fabian. *Confidence Measurement Techniques in Automatic Speech Recognition and Dialog Management*. Der Andere Verlag, 2008.

[3] D. Falavigna, R. Gretter, and G. Riccardi. Acoustic and word lattice based algorithms for confidence scores. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 1621-1624, 2002.

[4] L. Gang, C. Wei, and G. Jun. Novel active learning method for speech recognition. *China Communications*, vol. 7, no. 5:29–39, Nov. 2010.

[5] D. Hakkani-Tur, G. Riccardi, and A. Gorin. Active learning for automatic speech recognition. In *Acoustics, Speech, and Signal Processing (ICASSP). 2002 IEEE International Conference on*, volume 4, pages IV–3904. IEEE, 2002.

[6] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka. Speech modeling based on committee-based active learning. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4350–4353. IEEE, 2010.

[7] D. J. Hsu. Algorithms for active learning. 2010.

[8] H. Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

[9] T. M. Kamm. *Active Learning for acoustic speech recognition modelling*. PhD thesis, Citeseer, 2004.

[10] T. Nadungodage and R. Weerasinghe. Continuous sinhala speech recognizer. In *conference on Human Language Technology for Development, Alexandria, Egypt*, 2011.

[11] F. Olsson. A literature survey of active machine learning in the context of natural language processing. SICS Technical Report T2009:06, Swedish Institute of Computer Science, 2009.

[12] G. Riccardi and D. Hakkani-Tu¨ r. Active and unsupervised learning for automatic speech recognition. In *Proc. Eurospeech*, volume 2, page 3. Citeseer, 2003.

[13] G. Riccardi and D. Hakkani-Tur. Active learning: Theory and applications to automatic speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 13(4):504–511, 2005.

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

[15] S. Tong. *Active learning: theory and applications*. PhD thesis, Citeseer, 2001.

[16] G. Tur, R. E. Schapire, and D. Hakkani-Tur. Active learning for spoken language understanding. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–276. IEEE, 2003.

[17] F. Wessel. *Word posterior probabilities for large vocabulary continuous speech recognition*. PhD thesis, Universita¨tsbibliothek, 2002.

[18] F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 9(3):288–298, 2001.

[19] Y. Wu, R. Zhang, and A. Rudnicky. Data selection for speech recognition. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 562–565. IEEE, 2007.

[20] S. J. Young and S. Young. *The HTK hidden Markov model toolkit: Design and philosophy*. Citeseer, 1993.

[21] D. Yu, B. Varadarajan, L. Deng, and A. Acero. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech & Language*, 24(3):433–444, 2010.