N-BEST ENTROPY BASED DATA SELECTION FOR ACOUSTIC MODELING

Nobuyasu Itoh[⋆] *Tara N. Sainath*[†] *Dan Ning Jiang*[‡] *Jie Zhou*[‡] *Bhuvana Ramabhadran*[†]

[⋆] IBM Research - Tokyo, IBM Japan Ltd., Yamato, Japan 242-8502 iton@jp.ibm.com
[†] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA tsainath,bhuvana@us.ibm.com
[‡] IBM Research - China, #19A, Zhongguancun Software Park, Haidian District, Beijing, P.R.C. 100193, jiangdn,jiezhj@cn.ibm.com

ABSTRACT

This paper presents a strategy for efficiently selecting informative data from large corpora of untranscribed speech. Confidence-based selection methods (i.e., selecting utterances we are least confident about) have been a popular approach, though they only look at the top hypothesis when selecting utterances and tend to select outliers, therefore, not always improving overall recognition accuracy. Alternatively, we propose a method for selecting data looking at competing hypothesis by computing entropy of *N*-best hypothesis decoded by the baseline acoustic model. In addition we address the issue of outliers by calculating how representative a specific utterance is to all other unselected utterances via a *tf-idf* score. Experiments show that *N*-best entropy based selection (%relative 5.8 in 400-hour corpus) outperformed other conventional selection strategies; confidence based and lattice entropy based, and that *tf-idf* based representativeness improved the model further (%relative 6.2). A comparison with random selection is also presented. Finally model size impact is discussed.

Index Terms— *N*-best entropy, Acoustic modeling, Active learning, Data selection, Speech recognition

1. INTRODUCTION

In this paper we present a method for selecting a relevant subset of training data from a large data pool of untranscribed speech. The increase in amount of speech data from a variety of sources, the need for effective unsupervised training [1] is becoming more important. Transcriptions are usually expensive and time-consuming. It is not a good strategy to transcribe sufficient amount of speech data at random even if raw data is easily obtained, as in the case of call centers. The objective of data selection for acoustic modeling is to identify the relevant data, which improves the word error rate, and to reduce the number of training examples that must be transcribed [2]. Active learning approaches to these kinds of problems have a long history in machine learning research [3], and have been applied to various areas where statistical modeling techniques are used, such as natural language processing [4] and acoustic modeling. One of the common principles in active learning data selection is that samples for which the predictor or recognizer has the higher uncertainty are more informative for improving the current model. In speech recognition the *confidence score* is a common methodology for assessing this informativeness (i.e. [2, 5]). Recently, entropy has also been explored for calculating the uncertainty. Yu [6] proposed a unified framework for active learning and semi-supervised learning based on a global entropy-reduction principle. Hamanaka [7] proposed a method for selecting data with a committee-based approach. Another principle for active learning is that more frequently found (more representative) samples are more valuable for training. There are many samples with high informativeness, but with less representativeness. These two principles are sometimes contradictory. Many researchers addressed this problem. For example, Huang [8] provides a min-max framework for selecting utterances considering both the informativeness and representativeness. However, active learning in acoustic modeling is still challenging, more efficient criteria are required. The key contributions of this paper are:

- *N*-best entropy criterion to sufficiently reflect the uncertainty in the baseline model
- The use of a phone-based *tf-idf* measure as a representativeness metric
- Combined informativeness and representativeness measures for data selection

This paper is organized as follows. Section 2 describes our framework for active learning. Section 3 outlines the informativeness criteria used in acoustic modeling, address their problems, and propose the *N*-best entropy criterion. In Section 4 we consider *representativeness*, introduce the *tf-idf* vector representing an utterance using phone multi-grams, and define a metric to calculate the representativeness. Section 5 presents our experiments using a transcription task and compares the results with conventional selection methods and with random selection. Section 6 discusses the results and concludes the paper.

2. OVERVIEW

Figure 1 is an outline of our active learning framework. First, we create an initial model from a limited amount of previously transcribed speech corpus data. Next we decode all of the utterances in the data pool using the initial model, and obtain the *N*-best hypotheses for each utterance. Then we calculate two statistical values; the *N*-best entropy, which will

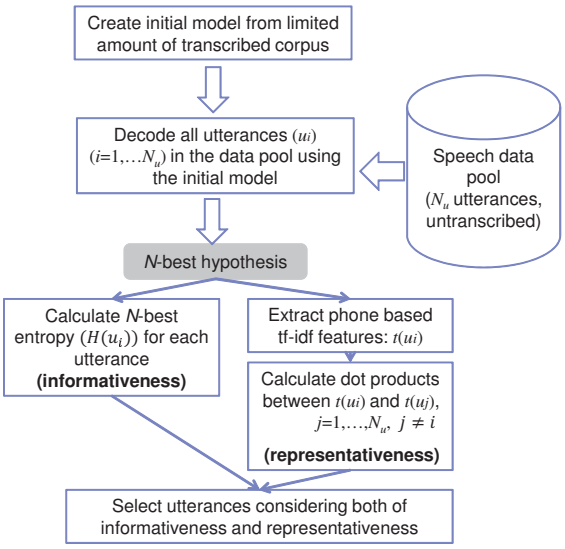


Fig. 1. Overview of our proposed method for active learning

be defined in Section 3, and the mean distance between each sample, expressed by a *tf-idf* vector, and all of the other samples in the pool. These values are metrics of the *informativeness* and *representativeness*, respectively. Then utterances are selected from the pool while considering both of these values. Finally the selected utterances will be transcribed and added to the training corpus.

3. INFORMATIVENESS CRITERIA

3.1. Previously proposed criteria

3.1.1. Confidence scores based on word confusion networks

The confidence score is important not only for active learning but also for many practical applications [9]. A frequently used method to obtain this score is to perform word consensus network decoding [10]. A word consensus network is a sequence of bundled words (called a *bin*), a group of word candidates aligned with their shared time intervals. A posterior probability is assigned to each alternative word in the *bin*. The probability assigned to the most possible word can be regarded as a *confidence score* (C_w). We therefore can define a utterance (u) level confidence score as

$$CS_u = \frac{\sum_{w \in \mathcal{W}} C_w T_w}{\sum_{w \in \mathcal{W}} T_w}, \tag{1}$$

where $\mathcal{W}$ is the set of the best candidates in the utterance u and T_w is the duration of word w [11]. CS_u is the average confidence score for utterance u. One problem with this selection criterion is that it uses only the best candidate to calculate scores, and so they might not sufficiently reflect the uncertainty. Additionally low-confidence utterances are frequently outliers in the distribution of the training samples, and they do not always contribute to improve the acoustic models.

3.1.2. Lattice entropy

The decoded result of an utterance can be represented in a word lattice structure, where a combinatorial number of hypotheses are present with their probabilities. Yu *et al.* [6] proposed a lattice entropy method for evaluating the uncertainty of utterances, which is calculated from a lattice generated by the decoder. If $\mathcal{L}_u$ is the set of all path in the lattice of the utterance u, then lattice entropy $H_u$is defined as

$$H_u^{Lattice} = -\sum_{q \in \mathcal{L}_u} p_q log(p_q), \tag{2}$$

where p_q designates the posterior probability of path q.

3.2. *N*-best Entropy

Lattice entropy should be affected by utterance length. More specifically longer utterances tend to have higher entropy. The sentence length balance for active learning has been discussed in the natural language processing community. Becker [12] pointed out that length-balanced sampling is an important factor for the effectiveness of active learning for a statistical parser. We therefore propose *N*-best entropy as a metric to evaluate the uncertainty of the decoded results. If $\mathcal{N}_u$ is an *N*-best list, then *N*-best entropy ($H(u)$) of an utterance (u) is calculated from the *N*-best hypotheses generated by the decoder. It is defined as

$$H_u^{N-best} = -\sum_{q \in \mathcal{N}_u} p_q log(p_q), \tag{3}$$

where p_q designates the posterior probability of the hypothesis q. Additionally we should re-normalize each p_q so that it satisfies $\sum_{q \in \mathcal{N}_u} p_q = 1$ before calculating H_u^{N-best}. Then we obtain the criterion which is not affected by utterance length.

4. REPRESENTATIVENESS

As we stated in Section 1, *representativeness* means how frequently a sample is found in the data pool. In other word it is a indicator of how much a given sample is relevant to the overall pool. We must therefore define a metric space where utterance samples can be compared to assess the representativeness. Yu [6] used the Kullback-Leibler divergence between the lattices to measure the distance between corresponding utterances, but it is time consuming to align all of the paths in lattices. In addition distance calculations that is proportional to the square of the number of utterances are required to estimate the representativeness of all of the samples in the pool. In our approach we use a phone-based *tf-idf* vector to characterize each utterance and estimate the distance in a simpler way. In this section we describe our features and our definition of distance.

The *tf-idf* is one of the representations for a document in a vector space model, and is often used in information retrieval. The term frequency (*tf*) is the normalized occurrence of each term in the document, and the inverse document frequency (*idf*) is a metric of the ”importance” of that term. Both *tf* and

idf are defined as

$$tf(term, d) = \frac{\text{the occurrence of } term}{\text{total number of terms in } d} \quad (4)$$

$$idf(term) = log\frac{|D|}{1 + |\{d : term \in d\}|}, \quad (5)$$

where $|D|$ is the total number of documents and $|\{d : t \in d\}|$ is the number of documents where the term t appears. Then the *tf-idf* value can be defined as the product of *tf*() and *idf*(). A document is now characterized with a single *tf-idf* vector, with each value corresponding to the *tf-idf* value of one word. If we consider each utterance as one document and each phone *n*-gram as one term, then we can calculate the *tf-idf* vector for each utterance, where the dimension of vectors is the number of distinct phone *n*-grams. The distance, the metric of similarity between two vectors ($\boldsymbol{t}(u_j)$, $\boldsymbol{t}(u_k)$), between utterance u_j and u_k can be defined in various ways. The similarity (sim) is typically defined as the direction cosine of the vectors.

$$sim(\boldsymbol{t}(u_j), \boldsymbol{t}(u_k)) = \frac{\boldsymbol{t}(u_j) \cdot \boldsymbol{t}(u_k)}{|\boldsymbol{t}(u_j)||\boldsymbol{t}(u_k)|} \quad (6)$$

Now we can define the representativeness of an utterance u_k in the data pool which includes N_u utterances by measuring the average similarity with all other utterances in the pool.

$$\theta_{u_k}^{REP} = \frac{1}{N_u} \sum_j sim(\boldsymbol{t}(u_j), \boldsymbol{t}(u_k)) \quad (7)$$

The process of creating phone *n*-grams to calculate the *tf-idf* vector for an utterance is descended as follows:

1. Create phone multi-grams from the initial training data with transcriptions and forced alignments:
   (a) List the phone *n*-grams ($n = 1, \ldots, L$) and count them, where cross-word-boundary *n*-grams are excluded.
   (b) Select the *n*-grams with a higher frequency than a threshold.
   (c) Segment the original phone sequences into selected *n*-grams with the longest matches, and filter out the unused *n*-grams for $n \geq 2$.
2. Segment all of the base-forms in the phonetic dictionary into the phone multi-grams with the longest matches.

As shown in Figure 1, we decode each utterance and obtain the *N*-best hypothesis. Then each hypothesis (word sequence) is converted into phone *n*-gram sequence by referring to the phonetic dictionary with the segmented base-forms. The *N*-best phone *n*-gram sequence is regarded as one document, and the *tf-idf* vector is calculated.

This leads to two criteria for our data selection for each utterance. We combine them with interpolation so that if the informativeness and representativeness of a utterance u are θ_u^{INF} and θ_u^{REP}, then the combined criterion of u is

$$\theta_u^{COMB} = (\theta_u^{INF}) \times (\theta_u^{REP})^{\lambda}. \quad (8)$$

The λ is a weighting coefficient, which should be set to maximize the accuracy of the development set [13].

5. EXPERIMENTS

5.1. Setup

We conducted experiments using our in-house speech corpus of voice mail transcriptions in a business domain in English. All of the experiments are based on speaker-independent models that are discriminatively trained with large vocabulary continuous speech recognition. We prepared two data pools and baseline models: One is a total of 1.6 K-hours of data (approximately 1.1M utterances) and the initial model was created from a separate 200 hours of data. The other set is 400 hours of data with 50 hours in the initial model. 400 hours of data and 50 hours of data are the subsets of the 1.6 K-hours pool and the 200 hours of data, respectively. Both acoustic models are phone-based with quinphone context-dependent states. Table 1 shows our model parameter settings. We used four test sets (s1, s2, s3, and s4) with different ranges of word error rates. They included approximately 32K, 32K, 9K, and 33K words. In the test set decoding, we built the vocabularies and 4-gram language models from the transcripts used for the initial models (50 hours and 200 hours), and interpolated the language models with a model built from a large telephone corpus. The vocabulary sizes were 17K (50 hours) and 37K (200 hours) respectively. We also used these LMs for word-confusion-network generation, which is required for calculating confidence scores [11]. We used 1-gram LM created from the same corpus for calculating the lattice entropy and the *N*-best entropy by considering the results of our preliminary experiment.

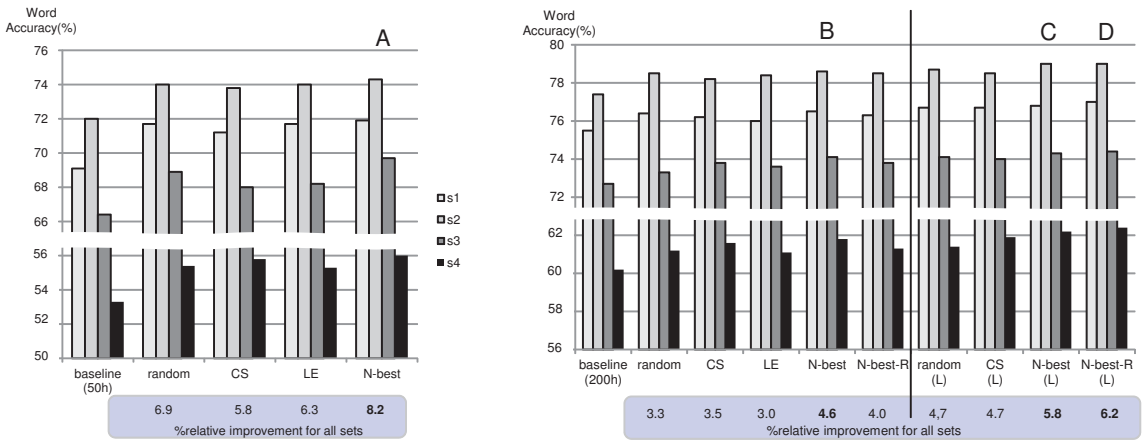
Table 1. Acoustic model parameters

	States	Gaussians
200 h small	5 K	150 K
200 h large	8 K	240 K
50 h	5 K	150 K

As described we conducted two experiments using different initial models and data pools. We selected 50 hours and 200 hours of data from the data pools of 400 hours and 1.6K hours, respectively, based on random, lowest confidence scores, highest lattice entropies, highest *N*-best entropies, and *N*-best entropies with the representativeness. Then we created 100-hour (50+50) models and 400-hour (200+200) models.

5.2. Results

The results are shown in Figure 2. The results of both cases (100-hour model, 400-hour model) are similar except for the absolute values of the accuracies. Confidence score (*CS*) is better for Set s4 in the 100-hour model and for Set s3 and s4 in the 400-hour model than random selection, which suggests that *CS* improves the accuracy of poorly recognized utterances, which is consistent with our intuition.



(a) 50+50 models

(b) 200+200 models

Fig. 2. Comparison of the data selection strategies, CS: Confidence score, LE: Lattice entropy, N-best: *N*-best entropy, N-best-R: *N*-best entropy with representativeness. (L) : *large* setup, Others: *small* setup

- Lattice entropy was not better than random selection in most cases.
- To the contrary, *N*-best entropy (**A, B** in Figure 2) achieved the highest gain over the baselines in every case.

The relative improvements for all of the sets were 8.2% (**A**) and 4.6% (**B**) in the 100-hour and 400-hour models respectively.

- The representativeness had a negative gain in the *small* setup of the 400-hour model, but improved the model further (**D**, %relative 6.2) in the *large* setup and outperformed not only CS and LE but also *N*-best entropy alone (**C**, %relative 5.8) in the configuration with the same parameters.

6. CONCLUDING REMARKS

We have described a framework for efficient data selection using the *N*-best hypotheses and their entropy. We compared our proposed method with random selection and two conventional selection strategies; the confidence score and the lattice entropy, and found that the *N*-best entropy with the representativeness was the best criterion tested.

7. REFERENCES

[1] S. Furui, "Generalization problem in asr acoustic model training and adaptation," in *ASRU*, 2009.

[2] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech reognition," in *ICASSP*, 2002.

[3] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active learning from stream data using optimal weight classifier ensemble," *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 40, pp. 1607–1621, 2010.

[4] J. Zhu, H. Wang, T. Yao, and B. K. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," in *COLING*, 2009.

[5] G. Riccardi and D. Hakkani-Tur, "Active learning theory and application to automatic speech recognition," *IEEE Trans. Speech Audio Process*, vol. 13, 2005.

[6] 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 and Language*, vol. 24, pp. 433–444, 2010.

[7] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," in *ICASSP*, 2010, pp. 4350–4353.

[8] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and repesentative examples," in *NIPS*, 2010.

[9] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, 2005.

[10] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, 2000.

[11] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for lvcsr," *Speech Communication*, vol. 52, pp. 652–663, 2010.

[12] M. Becker and M. Osborn, "A two-stage method for active learning of statistical grammars," in *IJCAI*, 2005.

[13] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *EMNLP*, 2008, pp. 1070–1079.