TRAINING DATA SELECTION BASED ON CONTEXT-DEPENDENT STATE MATCHING

Olivier Siohan

Google Inc., New York

ABSTRACT

In this paper we construct a data set for semi-supervised acoustic model training by selecting spoken utterances from a massive collection of anonymized Google Voice Search utterances. Semi-supervised training usually retains high-confidence utterances which are presumed to have an accurate hypothesized transcript, a necessary condition for successful training. Selecting high confidence utterances can however restrict the diversity of the resulting data set. We propose to introduce a constraint enforcing that the distribution of the context-dependent state symbols obtained by running forced alignment of the hypothesized transcript matches a reference distribution estimated from a curated development set. The quality of the obtained training set is illustrated on large scale Voice Search recognition experiments and outperforms random selection of high-confidence utterances.

Index Terms— data selection, semi-supervised training

1. INTRODUCTION

State-of-the-art speech recognition systems are typically trained in a supervised mode using manually labeled data. Unfortunately, transcribing a large amount of training data scales poorly when developing an application like Google Voice Search which is available in over 50 languages. For this reason, Google and others [1] rely more and more on semi-supervised training where an existing system trained from a small amount of labeled data is used to label a large data set extracted from the anonymized applications logs, often orders of magnitude larger than the manually labeled set.

A key aspect when training an acoustic model is the accuracy of the reference transcripts. In [1] it was shown that noisy reference labels are very detrimental when using modern discriminative training techniques such as boosted-MMI [2] or Deep Neural Networks (DNN). In an effort to improve the quality of the training labels, the authors proposed to transcribe logs data using multiple systems and system combination techniques. To validate the correctness of the hypothesized transcript, utterance-level confidence measures are used [3]. While high confidence is typically correlated with the accuracy of the recognition hypothesis, selecting only data with the highest confidence can be detrimental [4] as it limits the diversity of the data set and does not guarantee that the entire feature space will be sampled. In practice, it has been observed that a good strategy is to discard utterances with either a very low or a very high confidence [1].

If the unsupervised training set were randomly sampled from the logs, it would be an accurate characterization of the application data and very suitable for retraining a model (provided that the hypothesized transcripts are correct). The confidence-based sampling however can lead to a training set that no longer exhibits the same properties as the target application, as it may for example predominantly discard utterances spoken in a noisy environment or from non-native speakers which are harder to recognize.

This motivated us to propose in [5] an approach that attempts to match the distribution of the confidence-sampled data with the distribution of a target development set representing the application domain. To give a simple example, let us assume that the average signal-to-noise ratio (SNR) in the development set is 20dB. If the average SNR in the selected confidence-sampled set is 24dB, it may not be a very compelling training set since it differs significantly from the target set, at least in terms of SNR. It is then desirable to introduce a constraint in the data selection procedure so that the SNR of the selected set matches the SNR of the development set. This is the approach that we followed in [5] where the distribution of interest was the distribution of the iVector [6] associated to each utterance. The iVector of an utterance represents the coordinate of that utterance in a space characterizing the acoustic features. This resulted in a data selection procedure enforcing consistency between the acoustics of the selected data set and the target development set.

That work however failed to address the linguistic diversity of the selected data. For example, if the selected data contains an unusually large number of utterances with an identical hypothesized transcript, it would be a poor training set. Hence, in this paper we propose to use a similar distribution matching approach, but this time to enforce some phonetic consistency between the selected and target sets. The next section will describe the selection procedure based on a relative entropy criterion similar to [7]. It is followed by the description of the phonetic constraints, before presenting recognition results comparing the proposed data selection approach with random selection on a large scale Voice Search application.

2. DISTRIBUTION MATCHING FOR DATA SELECTION

2.1. Principle

Let $P(X)$ be the distribution of a random variable X characterizing utterances from the application domain and estimated from a development set. In this section, we will not define what X represents except that it is a random variable that can be obtained from an utterance. Let $Q(X)$ be the distribution of the same variable X, but this time estimated from a data set selected from the application logs. If the selected data is randomly sampled from the application logs, one should expect the distributions $P(X)$ and $Q(X)$ to be similar. In other words, the Kullback-Leibler divergence [8] $D_{KL}(P\|Q)$ between the 2 distributions should be close to 0.

However, as discussed in the Introduction, the selection approach is not random but biased by the requirement that the selected utterances are expected to have a high confidence to favor utterances with a correct recognition hypothesis. This biased selection will result in a distribution $Q(X)$ which may differ significantly from $P(X)$, leading to a sub-optimal training set for the application.

We propose to use the selection procedure originally described in [7] which iterates over a set of utterances from the logs and will add an utterance to the selected set only if it does not increase the

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

KL divergence $D_{KL}(P\|Q)$ between the reference distribution P and the distribution of the selected set Q. This is formally described in Algorithm 1 and will lead to the construction of a data set having a distribution Q close to P, based on the KL divergence.

Algorithm 1: Relative-entropy data selection algorithm

Input: A reference distribution P; an initial set of already selected utterances S; a large set of confidence-selected utterances U from the application logs

Output: The selected data set S

```
Estimate the distribution Q_S
D ← D_KL(P‖Q_S)
for each utterance u ∈ U do
    Estimate Q_{S∪u}
    D' ← D_KL(P‖Q_{S∪u})
    if D' < D then
        S ← S ∪ u
        D ← D'
return S
```

2.2. Context-Dependent State Distribution

In the previous section, we formulated the data selection problem in general terms without describing the distributions P and Q. Obviously, the nature of the algorithm imposes some practical constraints on the choice of those distributions. The first one is related to the derivation of $D_{KL}(P\|Q_{S\cup u})$. Because it has to be computed for each candidate utterance u, it has to be computationally efficient. Similarly, one should also be able to efficiently re-estimate the distribution $Q_{S\cup u}$ for each candidate utterance u.

In this work, we propose to characterize a data set by the distribution of the context-dependent (CD) Hidden Markov Model (HMM) state symbols obtained by aligning the utterance transcript (or hypothesized transcript) against the audio signal. The forced alignment is done following the standard procedure when training acoustic models. A word-level acceptor is constructed from the transcript with optional silences added at the beginning/end of the utterance as well as between words. The transcript acceptor is composed with a lexicon transducer, a context-dependency transducer and an HMM transducer to produce a forced-alignment decoding graph. Running Viterbi decoding then provides a sequence of context-dependent HMM state symbols along the alignment path. We propose to describe a set of utterances by the unigram distribution of the CD state symbols collected by running forced-alignment. Note that the data extracted from the logs may have been endpointed based on slightly different endpointer configurations since the Voice Search production engine is regularly updated. For that reason, the CD state symbols corresponding to the silence phone are discarded when estimating the distributions P and Q to prevent any skew related to variations in silence padding.

In Algorithm 1, the set of selected utterances S is initialized by randomly selecting a small set of utterances from the logs. Given that modern ASR systems typically operate on HMMs having an inventory of CD state symbols in the order of ten thousand states or more, the initial estimate of Q_S may not be very accurate when S is small. To alleviate this issue and similar to the approach followed in [7], we use the skew divergence [9], a smooth version of the KL divergence which interpolates Q_S using P:

$$D_{SD}(P\|Q_S) = \sum_{c\in\{\text{all CD states}\}} P(c) ln \frac{P(c)}{(1-\alpha)P(c) + \alpha Q_S(c)}$$

where c represents a CD-state index and α is a smoothing constant typically set in the range $[0.95 - 1]$. Note that when $\alpha = 1$, the skew divergence is equivalent to the KL divergence.

3. EXPERIMENTS AND RESULTS

3.1. Databases

All experiments are based on the Voice Search task and are reported on a variety of languages. For each language, we estimated the target unigram distribution P of the CD state symbols on a development set consisting of about 25 hours of mobile queries, mostly a mix of search queries and dictation-type sentences. The CD-state symbols corresponding to the silence phone were excluded from the distribution. We then extracted large data sets of high confidence utterances from the anonymized application logs either by using random sampling or by using the proposed CD-state matching constraint. In all experiments, the α constant used by the skew divergence was set to 0.95. As reported in [5], Algorithm 1 can converge quickly and then stops selecting utterances, preventing the construction of arbitrarily large sets. For that reason, we typically split the input set of high-confidence queries into multiple subsets and ran data selection independently in each subset. The selected sets are then merged, enabling the construction of data sets of arbitrary size. We typically aim at constructing training sets of about 3M utterances, which translates into 3,000 hours of training material. Depending on the target language, the size of the CD state inventory varied from about 5k to 15k tied-states. Depending on the experiments, the acoustic models were either HMM systems trained using several iterations of boosted-MMI or DNN models trained using our distributed DNN infrastructure [10]. In all DNN experiments, the network consisted of 7 fully connected hidden layers of 2560 neurons each that use RELU non-linearities [11] and was trained to minimize a cross-entropy loss using asynchronous gradient descent with a fixed learning rate of 0.003 and mini-batch updates of 200 frames.

3.2. Feedback loop oddities in unsupervised training

The procedure followed to build the Voice Search acoustic models relies heavily on automation, starting with the definition of the test vocabulary that is selected as the top N most-frequent words from typed queries, where N is in the order of a few million, depending on the target language. Similarly, the pronunciation lexicon is obtained mostly using a grapheme-to-phoneme (G2P) system. We observed that this procedure, combined with the frequent update of both the acoustic and language models trained from high-confidence utterances extracted from the logs can create a feedback loop that results in a word or short sentence being frequently hypothesized, especially when dealing with difficult queries such as queries spoken by children. For example, the left part of Table 1 represents the top-N queries from an unsupervised training set randomly sampled from high-confidence queries extracted from the Voice Search logs for British English (en-gb). This indicates that the word 'kdkdkdkdkdkdkdkd' which was in the tail of the vocabulary list and whose G2P pronunciation is 'k eI d i k eI d i...' ended-up being the most frequently hypothesized word in the en-gb Voice Search logs. This issue did occur in many of the languages where Voice Search was constructed and frequently updated using unsupervised training.

kdkdkdkdkdkdkdkd	hello
hello	facebook
facebook	google
google	hi
yes	how are you
kdkdkdkdkdkdkdkd kdkdkdkdkdkdkdkd	no
hi	youtube
how are you	yes
hello hello	thank you
weather	good morning
okay	cancel
kdkdkdkdkdkdkdkd kdkdkdkdkdkdkdkd kdkdkdkdkdkdkdkd	ebay
thank you	what
youtube	send text
no	com

Table 1. Top-15 queries in two unsupervised training sets sampled from the en-uk Voice Search logs. left: random sampling of top confidence queries. right: random sampling of top confidence queries with CD-state matching constraint.

We applied the proposed CD-state distribution matching approach to extract an unsupervised training set from high-confidence utterances extracted from the en-gb logs. The reference CD-state symbol distribution was constructed from a manually transcribed development set of 25 hours of Voice Search queries randomly extracted from the anonymized application logs. The right part of Table 1 lists the top-N queries from the data set constructed again from high-confidence queries but under the CD-state distribution matching constraint. The word 'kdkdkdkdkdkdkdkd' is no longer one of the most frequent queries in the selected data set, which results in an unsupervised training set that better matches the seed distribution. The proposed data selection procedure is language agnostic and does not require any input resource other than what is already available, namely a test lexicon in order to run forced alignment against the hypothesized transcript.

3.3. CD-state distribution matching experiments

The first set of experiments was conducted using a standard HMM-based system in British English. The development set used to construct the reference CD-state distribution was a 25-hour set called MOBILE_20120901_20121031 consisting of a mix of search and dictation queries. We constructed two unsupervised training sets of 3.5M utterances (about 3,500 hours of audio), the first one randomly sampled from high-confidence logs data, the second one randomly sampled from high-confidence data and subject to the CD-state distribution matching constraint. For each training set, we built an HMM-based system estimated using boosted-MMI. Evaluations were carried out on the development set as well as two additional 25-hour long test sets named T2011_EN_GB_INTENT and T2011_EN_GB_VS corresponding to dictation and voice search utterances, respectively. Results are given in Table 2 in terms of word error rate (WER). The system trained on the high-confidence data set selected using the CD state distribution matching constraint outperforms the system trained on randomly selected data, leading to 1.3% to 2.2% absolute reduction of the WER. Note that while the constraint forced the training set to match the CD-state symbol distribution of the MOBILE_20120901_20121031 development set, it led to performance improvement on the unseen test sets as well.

The same 3.5M utterances training sets were used to build two DNN-based systems. Results are given in Table 3 and illustrate the effectiveness of the proposed approach which led to 0.4% absolute reduction in WER. Note that given the size of the test sets, a 0.1% reduction of the WER is statistically significant.

Baseline (Random Selection)	WER (ins/del/sub)
MOBILE_20120901_20121031	31.0 (4.8/7.3/18.8)
T2011_EN_GB_INTENT	34.8 (6.5/7.5/20.8)
T2011_EN_GB_VS	47.0 (6.6/11.8/28.5)
Proposed (CD-state matching Selection)	**WER (ins/del/sub)**
MOBILE_20120901_20121031	29.3 (4.8/6.3/18.2)
T2011_EN_GB_INTENT	33.4 (6.6/6.6/20.3)
T2011_EN_GB_VS	45.4 (6.8/10.7/28.0)

Table 2. HMM-based systems (en-gb) trained on randomly selected high-confidence data (Baseline) and high-confidence data subject to CD-state distribution matching constraint (Proposed).

Baseline (Random Selection)	WER (ins/del/sub)
MOBILE_20120901_20121031	24.9 (4.6/5.3/15.0)
Proposed (CD-state matching Selection)	**WER (ins/del/sub)**
MOBILE_20120901_20121031	24.5 (4.7/5.1/14.8)

Table 3. DNN-based systems (en-gb) trained on randomly selected high-confidence data (Baseline) and high-confidence data subject to CD-state distribution matching constraint (Proposed).

Experiments were also run on Indian English (en-in). A 25-hour long data set named MOBILE_20130101_20130430 was used to estimate the reference distribution P. Two data sets of 2.1M high-confidence utterances (about 2,100 hours) were constructed: the first one, set (a), randomly selected, the second one, set (b), selected using the CD-state matching distribution constraint. DNN-based systems were trained from those two sets and the results are given in Figure 1 as a function of the number of stochastic gradient descent (SGD) updates during training. The data set constructed based on the CD-state matching constraint outperforms a randomly selected training set of equivalent size.

3.4. Triphone distribution matching experiments

In the next series of experiments, instead of characterizing a data set by the distribution of its CD-state symbols, we used its triphone symbol distribution. For each utterance, the sequence of triphone symbols was obtained by running forced alignment of the hypothesized transcript. As in the previous section we constructed a set of 2.1M high-confidence utterances, this time using the triphone distribution matching constraint. The results in Figure 1 illustrate that the data set (c) based on the triphone distribution matching outperforms set (b) based on the CD-state distribution. The systems built on the randomly selected data set (a) and the triphone-based data set (c) were also evaluated on a separate 25-hour test set, T2011_EN_IN_VS, distinct from the one used to estimate the reference distribution P. Results are given in Figure 2 and illustrate the effectiveness of the proposed data selection approach over random selection.

4. CONCLUSIONS

We proposed an algorithm to construct a training set from a large pool of Voice Search utterances. The data selection approach enforces that the distribution of the CD-state symbols obtained by running forced alignment on the selected utterances should match a reference CD state symbol distribution estimated from a development set. Note that instead of characterizing the data set by the CD state symbol distribution, it is also possible to use its triphone symbols distribution, which leads to improved performance. The approach

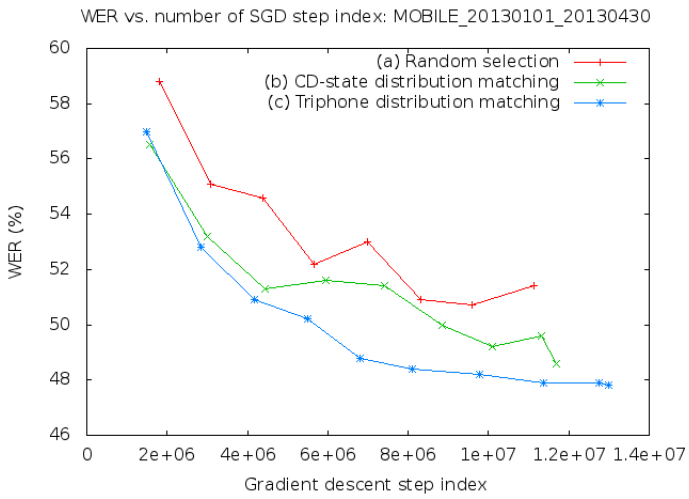

Fig. 1. WER as a function of the number of SGD steps for en-in on MOBILE_20130101_20130430. Three systems constructed from training sets of similar size: (a) one based on randomly selected high-confidence data, (b) one based on the CD-state distribution matching constraint, (c) one based on the triphone distribution matching constraint.

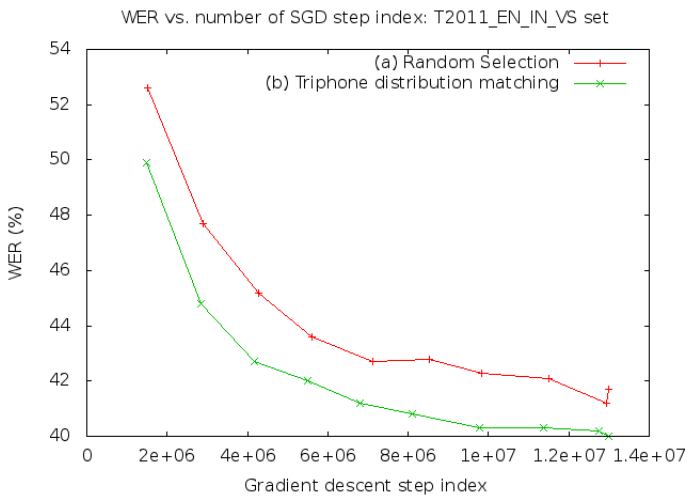

Fig. 2. WER as a function of the number of SGD steps for en-in on T2011_EN_IN_VS. Two systems constructed from training sets of similar size: (a) one randomly selected, (c) one based on the triphone distribution matching constraint.

was also shown to be effective in discarding some of the artifacts related to the feedback loop caused by frequent model updates in unsupervised training. This led to unsupervised training sets which better matched the testing conditions and improved recognition performance.

5. REFERENCES

[1] Yan Huang, Dong Yu, Yifan Gong, and Chaojun Liu, "Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, 2013.

[2] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George A. Saon, and Karthik Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[3] Hui Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[4] Rong Zhang and Alexander I. Rudnicky, "A new data selection approach for semi-supervised acoustic modeling," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[5] Olivier Siohan and Michiel Bacchiani, "iVector-based acoustic data selection," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, 2013.

[6] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[7] Abhinav Sethy, Panayiotis G. Georgiou, Bhuvana Ramabhadran, and Shrikanth S. Narayanan, "An iterative relative entropy minimization-based data selection approach for n-gram model adaptation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 1, pp. 13–23, 2009.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[9] Lillian Lee, "Measures of distributional similarity," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, ACL, pp. 25–32.

[10] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato, Andrew W. Senior, Paul A. Tucker, Ke Yang, and Andrew Y. Ng, "Large scale distributed deep networks.," in *NIPS*, 2012, pp. 1232–1240.

[11] M.D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q.V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G.E. Hinton, "On rectified linear units for speech processing," in *38th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, 2013.