Kullback-Leibler divergence-based ASR training data selection

Evandro Gouvêa
European Media Laboratory GmbH, Heidelberg, Germany
evandro.gouvea@alumni.cmu.edu

Marelie H. Davel
Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
marelie.davel@gmail.com

Abstract

Data preparation and selection affects systems in a wide range of complexities. A system built for a resource-rich language may be so large as to include borrowed languages. A system built for a resource-scarce language may be affected by how carefully the training data is selected and produced.

Accuracy is affected by the presence of enough samples of qualitatively relevant information. We propose a method using the Kullback-Leibler divergence to solve two problems related to data preparation: the ordering of alternate pronunciations in a lexicon, and the selection of transcription data. In both cases, we want to guarantee that a particular distribution of n-grams is achieved. In the case of lexicon design, we want to ascertain that phones will be present often enough. In the case of training data selection for scarcely resourced languages, we want to make sure that some n-grams are better represented than others. Our proposed technique yields encouraging results.

Index Terms: acoustic model training, lexical model, maximum entropy, Kullback-Leibler divergence, training data selection

1. Introduction

Data selection affects accuracy of automatic speech recognition (ASR) systems of varying degrees of complexity. An ASR system’s accuracy is affected by how well its knowledge sources can model the unseen test data. The knowledge sources, i.e. the acoustic, pronunciation, and language models, are trained from data that, if not carefully selected, will not have enough samples of important “events”. This lack of relevant data, in turn, results in models that perform poorly.

In the case of acoustic models, data selection affects large and small systems. Designers of large vocabulary continuous speech recognition (LVCSR) systems in resource-rich languages normally use all available data. However, the triphone occurrences differ sharply between text corpora [1]. Therefore, training models with all available data may result in acoustic models that are far from representative of a domain poorly represented in the training data.

Designers of ASR systems in resource-scarce languages, on the other hand, need to be careful about how to pick the data used for training from the small amount available. As Barnard [2] points out: “The optimal data distribution (for ASR training) is not exactly the same as the natural data distribution of phones / triphones etc. This makes intuitive sense: highly frequent units do not add much to accuracy after a certain point, and very rare units have little impact on test scores – so it is the middle range that needs to be boosted.”

Consider now the design of the pronunciation lexicon, from now on referred to simply as the dictionary. In a scenario where the dictionary is built for a target language, but includes words borrowed from other languages, it makes sense, from a human point of view, that the pronunciations using the phone set of the target language appear first. But if the dictionary also includes a pronunciation using a phone set from the original language (in case the system has to handle speakers who know how to pronounce the word in the original language), then the phones from languages other than the target language will never appear as the first alternate pronunciation in the dictionary.

ASR systems’ trainers start by linearly assigning audio frames to Hidden Markov Model (HMM) states. This initial assignment is used by the system to estimate non-flat (non-uniform) HMMs, which, by iteration, become increasingly more specific and accurate. For the initial segmentation, however, the trainers have no information that would allow it to choose one particular pronunciation of a word rather than other, in cases where a word has multiple pronunciations. Trainers normally choose the first alternate for this initial segmentation. If a phone never appears in the first alternate pronunciation of a word, as described in the scenario above, the trainer will never see it, and the model for that phone will not be initialized. This is a seemingly simple but practical problem that normally causes the trainer to break in subsequent steps.

Data selection can also be applied to the language model. Using all available data is not always the best approach [3], as relevant information changes from one domain to another. In this work, we do not examine data selection for language models, but we report on selection for acoustic model training and dictionary organization. The theoretical background for our method is presented in Section 2. In Section 3 we report on our work on automatically reorganizing a dictionary. In Section 4 we present our work on data selection for acoustic model training. We conclude in Section 5.

2. Background and Previous Work

We tackle two problems that, although seemingly different, can both be described as selecting data conforming to a predefined phone distribution, or, more generally, n-gram distributions.

The first problem concerns organization of a dictionary containing words with multiple pronunciations. The initial step of the trainer uses the first of these alternates, referred to as the first pronunciation in the remainder of this paper. Our goal is to reorder the alternate pronunciations associated with each word in a way that every phoneme in the dictionary appears at least once in a first pronunciation. Let’s consider the phone distribution $P_F$ of the phones appearing in the first pronunciation only. Our goal is then to reorganize the dictionary, i.e. reorder the alternate pronunciations, so less frequent phones in $P_F$ are favored. In other words, we want $P_F$ to approach the uniform
distribution. The second problem concerns selection of data to train acoustic models. Our goal here is to obtain a training set where the distribution of n-grams is as close as possible to a predefined one, possibly not a uniform distribution.

The Kullback-Leibler (KL) divergence [4] is often used to measure the distance between distributions. Two discrete random variables $X$ and $Y$ with distributions $p$ and $q$, respectively, yield the KL divergence as in Equation 1.

$$D_{\text{KL}}(X||Y) = \sum_{i} p(x_i) \log \frac{p(x_i)}{q(y_i)}$$

(1)

Entropy is a measure of the uncertainty of a random variable. It is a common way of measuring the distance between a distribution and the uniform distribution. A discrete random variable $X$ taking values $x_1, x_2, \ldots, x_n$ will have an entropy defined as in Equation 2.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

(2)

Notice that the entropy of the distribution $p(x)$ is equivalent to the KL divergence between the distribution $p(x)$ and a uniform distribution, i.e. the KL divergence if the distribution $q(y)$ is uniform, other than some constant and the sign. The entropy is maximized when all $x_i$ are equiprobable, i.e. the random variable $X$ is drawn from a uniform distribution. Likewise, the KL divergence is minimized when the distributions $p(\cdot)$ and $q(\cdot)$ are identical.

Entropy has been used to select data for acoustic model training [5] or to build language models [3]. Both of these focus on word distribution targeting a uniform distribution. Our work focuses on distributions of phones or n-grams rather than words. We use the entropy criterion for the task of reorganizing the dictionary. Selection of data on distributions for an ASR system uses KL divergence, as we have a target distribution. In this case, we noticed that the resulting selection tended to remove n-grams that were not very frequent, consistent with the criterion of minimizing divergence. To alleviate this issue, we use a criterion motivated by the minimum description length or Bayesian information criterion (BIC), given by Equation 3:

$$D = D_{\text{KL}}(X||Y) + \alpha \cdot \log(N)$$

(3)

where $N$ is the number of n-grams with non-zero count and $\alpha$ is an empirical constant used to balance the dynamic range of the two components on the right hand side.

3. Dictionary organization

The acoustic model trainer initializes the HMMs by linearly aligning audio to words in the training data. If a word has multiple pronunciations, most systems simply pick the first pronunciation for this initial alignment. If a phone never appears in any of the first pronunciations, its model is never initialized. This situation may occur, e.g. if a system is built for a target language, but the dictionary also includes borrowed words with phones from phone sets not in the target language. It could also happen if the dictionary was created by a grapheme to phoneme converter that can hypothesize multiple pronunciations per word [6], without human post-processing.

We propose a systematic, automated way of reordering pronunciations. This reordering may be crucial to allow the trainer to go beyond its initialization. (Further steps can use forced alignment to select which alternate pronunciation to use during later processing.) The reordering is based on a maximum entropy criterion, as described in Section 2.

3.1. Dictionary Construction

We formulate our entropy based sorting method as follows. Given a dictionary, we initialize the phoneme distribution $P_F$ from the pronunciations of words that have a single pronunciation. This distribution can take into account the words in the dictionary only, or it can consider the frequencies in which these words appear in a particular transcription set.

We then traverse the set of words that have multiple pronunciations. We update the $P_F$ with counts of phone that appear as first pronunciation only. Therefore, choosing one of the alternate pronunciations as a first pronunciation has the effect of changing the $P_F$. We choose as first pronunciation the alternate that maximizes the entropy, i.e. the alternate that most makes $P_F$ approach the uniform distributions.

3.2. Experimental Setup

We compare the accuracy obtained with models trained with dictionaries reorganized so that the first pronunciation is one of the four schemes: first alternate, last alternate, random, and maximum entropy-based. This comparison gives us a sense of whether there is any difference in performance. The choice of which scheme to use would then be dictated by whether the initialization works or breaks.

As a platform to test our dictionary reorganization schemes, we used the open source RWTH Aachen speech recognition system [7]. The acoustic models used by the decoder are triphone HMMs, trained from around 35 hours of narrowband speech. The incoming features are standard Mel Frequency Cepstral Coefficients (MFCCs). The language model is a 4-gram with Kneser-Ney discounting containing around 400,000 unigrams. The test set has around 3.1 hours of telephone quality data. Recognition was performed without resorting to state of the art techniques such as speaker adaptation or discriminative training, as our goal was to compare the relative merits of dictionary reorganization techniques. Therefore, the baseline performance is modest, reflecting the difficulty of the test.

3.3. Results and Discussion

The practical goal of automatically reorganizing the dictionary is to avoid the situations where the trainer initialization fails, while keeping the accuracy. One possible drawback of the maximum entropy-based scheme is that we may be forcing the wrong pronunciation for initialization. Consider the case e.g. of imported words. When we impose the alternate with the “foreign” phones as the first pronunciation, we force the trainer to initialize the “foreign” phones. However, the audio sample may be completely inappropriate as initialization for these phones. Moreover, if we contrast this with a random choice, we may be favoring pronunciations that are rare in the language, which a random sort would spread over several words.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First alternate pronunciation</td>
<td>N/A</td>
</tr>
<tr>
<td>Hand edited first alternate</td>
<td>32.8</td>
</tr>
<tr>
<td>Last alternate pronunciation</td>
<td>37.7</td>
</tr>
<tr>
<td>Random alternate pronunciation</td>
<td>36.9</td>
</tr>
<tr>
<td>Maximum entropy opt</td>
<td>36.9</td>
</tr>
</tbody>
</table>

Table 1: Word accuracy for the different lexical model reorganization schemes. The construct name refers to how the trainer initialization chose which alternate pronunciation to use. Accuracy of N/A means that the initialization failed.
Table 1 presents the speech recognition results obtained with acoustic models initialized according to the schemes presented in Section 3.2. Results show that, despite the possible drawbacks, the maximum entropy method satisfactorily reordered our dictionary without jeopardizing the system’s accuracy. At the same time, successful trainer initialization can be guaranteed with this method to an extent impossible to do with random selection.

4. ASR Training Data Selection

In resource-scarce environments, the most typical way of developing a new ASR corpus (or extending an existing corpus) is through targeted speech collection using tools such as Data-hound [8] or Woetzela [9]. Text prompts are then generated prior to recording, and the matching audio captured from multiple speakers. As only a limited set of recordings can be obtained this way (relative to the number of recordings available in resource-rich environments), the subset of transcriptions that are selected for recording should be chosen with care. Given a set of candidates for recording, it is not obvious what the optimal subset will be in order to obtain reliable acoustic models.

Our goal in selecting data for training is to boost units that are not already very frequent, and not very rare. Units could be phones, but more generically, n-grams. Given this goal, selecting data based on maximum entropy is not sufficient, as this would favor a uniform n-gram distribution, boosting rare n-grams more. KL divergence, on the other hand, allows us to choose a target distribution, which the subset will approach.

The data selection is determined as follows. We select the number of utterances $N$ that we want to extract from the training data $T$. We initialize the candidate subset of selected utterances $S$ to the first $N$ utterances from $T$ and compute the distribution $P_S$ of n-grams in $S$. For each remaining utterance $U$ in $T$, we compare $U$ with each utterance in $S$ and compute the change in KL divergence that would occur if we replaced the utterance with $U$. We then find $V$ in $S$ for which the KL divergence decreases most, and replace $V$ with $U$. We continue until the end of $T$. The final $S$ is extracted. The KL divergence is computed relative to a user provided distribution (which could also be the uniform distribution).

In our experiments below we simulate data selection for a resource-scarce language by training with a limited subset of the Wall Street Journal (WSJ) database [10]. We define the two application domains involved as (1) the target domain of the specific ASR application (consisting of all possible utterances the ASR application should be able to recognize) and (2) the source domain consisting of all the possible text candidates from which we would like to select a set of prompts for recording purposes. The source domain would typically be more general, while the target domain may be very specific. We experiment with selecting a subset of data from the source domain in different ways and train acoustic models using these subsets. We then evaluate the effectiveness of the different acoustic models when used for recognition in the target domain.

4.1. Speech Recognition System Setup

We use the speaker independent portion of the WSJ0 database to train acoustic models. We develop a fairly standard HMM-based recognizer using the HMM ToolKit (HTK) [11] for our recognition experiments. The system is built with conventional MFCCs concatenated with deltas and delta-deltas. Cepstral Mean Normalisation (CMN) is applied to the 39-dimensional features. We build HMMs with cross-word tied-state triphones, where state output probability density functions have 8-mixture multivariate Gaussian components with a diagonal covariance matrices, and apply a 40 class semi-tied transform. As we want to isolate the effect of data selection on the acoustic model only, and not on the language model, we use a simple flat phone loop grammar and perform phone recognition.

4.2. Data setup

Our goal is to simulate data selection for a resource-constrained language using the WSJ0 database. As such, we first partition the WSJ0 into a separate training and test set. From the test set we select different subsets of utterances that each represent a different “target domain” (see above). In order to represent different target domains, we randomly select a few utterances, and then add utterances with similar n-gram distributions until a 500-utterance test set is obtained. These target domains therefore have a different distribution from the main corpus, but still represent realistic ASR domains.

We now use the training set as our “source domain” and select 1000-utterance training subsets from this source domain, in such a way that the distribution follows one of the three main criteria: selection matching a target distribution, selection matching a uniform distribution, and random selection. Another dimension we explore is the order of the n-gram used to compute the KL divergence. Although the phone (unigram) distribution is more reliable than an estimate of the trigram distribution because of the sparsity of data, we believe the phone context to be more important to the robustness of the resulting acoustic models than simply considering the number of observations of each phone.

Specifically we create the following subsets:

- The target subset is selected by matching the subset’s n-gram distribution to the target distribution by minimizing the KL divergence. Two different subsets are obtained: one where unigrams are used during selection, and another where trigrams are used.
- The uniform subset is selected by matching the subset’s distribution to the uniform distribution by maximizing the entropy. This approach, as used in [5], is based on the intuition that this criterion balances the number of observations of each n-gram, increasing the less frequent ones, and decreasing the most frequent ones. Again, two different subsets are obtained, based on selection using unigrams or trigrams.
- The random subset is selected as a control experiment. The random sample follows the distribution of n-grams from the source domain. Therefore this subset allows us to create an acoustic model comparable to the other ones, as the amount of data is similar, and still be a representation of the original data.

We take care to ensure the acoustic content in the different data sets is similar. As some criteria tend to favor longer words over shorter words (when selecting on trigrams rather than unigrams) or shorter sentences over longer sentences (when limiting divergence rather than random selection), just selecting a specific number of utterances will result in some subsets having bad acoustic data. We therefore select approximately 1000 utterances per subset, but ensure that the number of phonemes are similar (approximately 64 200 in each case).
Table 2: Symmetric KL divergence for the different training data selection criteria, computed on the resulting subset using unigrams or trigrams. A divergence closer to zero is a better match to the target distribution; the target_test subset has a distribution approaching the distribution of n-grams in the test set.

<table>
<thead>
<tr>
<th>Subset</th>
<th>n-gram for selection</th>
<th>n-gram for KL calculation</th>
<th>1</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>1</td>
<td>0.04063</td>
<td>0.43399</td>
<td></td>
</tr>
<tr>
<td>uniform</td>
<td>3</td>
<td>0.02539</td>
<td>0.21365</td>
<td></td>
</tr>
<tr>
<td>random</td>
<td>N/A</td>
<td>0.01731</td>
<td>0.16200</td>
<td></td>
</tr>
<tr>
<td>target_test</td>
<td>1</td>
<td>0.00000</td>
<td>0.08511</td>
<td></td>
</tr>
<tr>
<td>target_test</td>
<td>3</td>
<td>0.00028</td>
<td>0.01670</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Results and Discussion

The key decision about which order of n-grams to use can be assessed by computing the resulting KL divergence from the selected subset. In fact, as we are using the KL divergence to select subsets of data, the KL divergence computed from the subsets assures us that the resulting divergence is indeed consistent with our criteria. Table 2 shows the symmetric KL divergence computed for subsets containing 1000 utterances, using unigrams and trigrams. (The symmetric KL divergence is \( D_{KL}(X||Y) + D_{KL}(Y||X)/2 \).) Notice that we can use any order of n-gram to select subsets, and independently compute the resulting KL divergence using any order of n-gram. The range of values of the KL divergence computed using different n-gram orders (i.e., n-grams with different \( n \), is different since the number of bins is different). Therefore, only comparisons on the same column of Table 2 are meaningful.

Most importantly, the KL divergence presented in Table 2 clearly shows that our data selection algorithm does select a set that has a distribution close to the intended distribution. Moreover, using trigrams to select subsets clearly results in subsets that are closer to the intended distribution than using only unigrams. If we use unigrams to compute the final KL divergence, subsets obtained by minimizing the KL divergence computed from trigrams are still much closer to the intended distribution in the case of the uniform subset, and close enough in the target_test case. Moreover, in all cases, the KL divergence of the target_test subset is better than the alternatives.

Table 3: Phone recognition results for the different training data selection criteria. The target_test and uniform sets have variants where the KL divergence was computed from either unigram or trigram distributions.

<table>
<thead>
<tr>
<th>Subset</th>
<th>n-gram for selection</th>
<th>Accuracy (%)</th>
<th>Correctness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uniform</td>
<td>1</td>
<td>58.19</td>
<td>66.45</td>
</tr>
<tr>
<td>uniform</td>
<td>3</td>
<td>61.31</td>
<td>69.26</td>
</tr>
<tr>
<td>random</td>
<td>N/A</td>
<td>62.39</td>
<td>70.80</td>
</tr>
<tr>
<td>target_test</td>
<td>1</td>
<td>63.88</td>
<td>72.86</td>
</tr>
<tr>
<td>target_test</td>
<td>3</td>
<td>66.26</td>
<td>73.71</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper we presented a method for the selection of data using Kullback-Leibler divergence. The criterion is used to select data so that the distribution of n-grams matches a desired distribution. This method was used for two separate, very practical problems in data preparation for ASR training: the design of a dictionary, and the selection of training data.

We have shown that a dictionary can be automatically re-ranged so that model initialization of the acoustic model trainer always finds at least some samples containing each phone.

We have also shown that a distribution targeted at the test set n-gram distribution, or, in practical terms, n-gram distribution estimated from the target application, is a better goal for data selection than the other more intuitive alternatives, namely a uniform distribution of n-grams or a distribution similar to the training data. Moreover, we have shown that using trigrams for such data selection is better than using unigrams, despite the possibility that the trigram distribution may be scarce.

6. References