DISCRIMINATIVE TRAINING METHODS FOR LANGUAGE MODELS USING CONDITIONAL ENTROPY CRITERIA

*Jui-Ting Huang**

University of Illinois at Urbana-Champaign
Electrical and Computer Engineering Department
Urbana, Illinois, USA

Xiao Li, Alex Acero

Microsoft Research
One Microsoft Way
Redmond, Washington, USA

ABSTRACT

This paper addresses the problem of discriminative training of language models that does not require any transcribed acoustic data. We propose to minimize the conditional entropy of word sequences given phone sequences, and present two settings in which this criterion can be applied. In an inductive learning setting, the phonetic/acoustic confusability information is given by a general phone error model. A transductive approach, in contrast, obtains that information by running a speech recognizer on test-set acoustics, with the goal of optimizing the test-set performance. Experiments show significant recognition accuracy improvements in both rescoring and first-pass decoding experiments using the transductive approach, and mixed results using the inductive approach.

Index Terms— Discriminative training, language model, unsupervised training, conditional entropy.

1. INTRODUCTION

A statistical language model (LM) is often trained in the maximum likelihood sense with some smoothing techniques, aiming at reducing the perplexity on future data. When used in speech recognition, however, such an objective may not be optimal. There have been a number of discriminative training criteria, including minimum classification error (MCE) [1], minimum word error (MWE) [2], large margin [3], and maximum conditional likelihood (MCL) [3, 4], which take into account acoustic confusability in language modeling. These methods are usually conducted in a supervised machine learning setting. In other words, both acoustic waveforms and their corresponding transcriptions are available at training time. Moreover, a speech recognizer is often used to produce a set of hypotheses (*e.g.*, an n-best list) for each train-set utterance. The LM parameters, then, are estimated to boost the likelihood of the correct hypothesis and to penalize those of the competing, incorrect hypotheses.

While these discriminative training methods lead to modest error rate reductions in different applications, they all require supervised (or implicitly supervised [4]) training data. This creates roadblocks to the development of speech applications in new domains where little in-domain acoustic data is available. For example, as an extension to voice search systems that recognize *local search* queries spoken by users [5], a *universal voice search* (UVS) system would allow users to speak general *web search* queries. Although we can leverage a large query log in the web search domain for language modeling, discriminative training of such an LM is challenging due to the lack of real acoustic data.

*The author conducted this work during her internship at Microsoft Research.

This work addresses the problem of discriminative training of LMs **without transcribed acoustic data**. We assume the availability of (1) an initial LM estimated from a text corpus, (2) a lexicon that contains pronunciations for all words in the LM, and (3) some representation of phonetic/acoustic confusability. We propose a new optimization objective that minimizes the conditional entropy of word sequences given phone sequences. Following this criterion, we explore two settings which differ in how phonetic/acoustic confusability information is obtained and what LM parameters are to be updated. In the first setting, we utilize a transducer to generate phonetically similar word sequences for each n-gram in the LM, and update n-gram probabilities under the conditional entropy criterion. This setting is akin to a recent work by Kurata et al. [6] that generated pseudo-ASR n-best lists for a sampled set of training sentences and applied MCE training accordingly. We will discuss the relation of our method to [6] in Section 3.2. The second setting, in contrast, runs a speech recognizer on test-set acoustic waveforms and generates acoustic confusability information therefrom. The discriminative training only updates LM parameters that occur in the n-best lists of the test set. In this regard, this setting essentially corresponds to an adaptation scenario: the discriminative training aims at optimizing the test-set performance and is heavily influenced by test-set inputs. To avoid confusion, we refer to these two settings as an *inductive* and *transductive* approach respectively. Finally, we evaluate both approaches on a UVS dataset.

2. CONDITIONAL ENTROPY CRITERIA

This section presents a discriminative training criterion for LMs that does not need transcribed acoustic data. We start by introducing our notation and problem setting. Let W and Φ denote random variables that represent word sequences and phone sequences respectively; and let their lower-case counterparts, w and ϕ, denote specific values of these random variables. Furthermore, we assume the availability of (1) an initial LM $\tilde{p}(w)$ trained in the maximum likelihood sense on a text corpus, (2) a pronunciation lexicon $\tilde{p}(\phi|w)$, and (3) some representation of phonetic (or acoustic) confusability information which will be discussed shortly. Given the above resources, our goal is to generate a new LM that minimizes the conditional entropy $H(W|\Phi)$ which is given by

$$\begin{aligned} H\left(W|\Phi\right) &= -\sum_{w}\sum_{\phi} p\left(w,\phi\right)\log p\left(w|\phi\right) \\ &= -\sum_{w}\sum_{\phi} p\left(w\right)p\left(\phi|w\right)\log p\left(w|\phi\right). \end{aligned} \tag{1}$$

The conditional entropy was used by Zweig and Nedel [7] as

a measure of how much information the phones provide about the words, and was shown in their work to correlate well with the phone-to-word transduction error rate for difference languages. This motivates us to use the conditional entropy as a training criterion for language modeling. Intuitively, minimizing the conditional entropy results in better predictability of words given phones.

Following [7], we re-write the conditional entropy as

$$H\left(W|\Phi\right) \approx -\sum_{w}\sum_{\phi}\tilde{p}\left(w\right)\tilde{p}\left(\phi|w\right)\log p\left(w|\phi\right). \quad (2)$$

Here we approximate $p(w)$ by the initial LM $\tilde{p}(w)$. In the universal voice search task, for example, $\tilde{p}(w)$ can be an n-gram LM estimated from a large set of text queries. In such a case, $\tilde{p}(w)$ represents the empirical distribution of w w.r.t. the text query corpus. Furthermore, while $p(\phi|w)$ represents the true distribution of all possible pronunciations given a word sequence, we approximate it using a typical lexicon $\tilde{p}(\phi|w)$ for simplicity. Finally, the posterior $p(w|\phi)$ is computed based on the Bayes rule,

$$p\left(w|\phi\right) = \frac{p_\lambda(w)^\alpha p(\phi|w)}{\sum_{w'} p_\lambda(w')^\alpha p(\phi|w')}, \quad (3)$$

where λ represents LM parameters (log n-gram probabilities in our case) that are to be updated. The sum in the denominator is taken over word sequences w' that are confusable with w. And $p(\phi|w)$ can be chosen to encode various forms of phonetic confusability. In one approach, we can obtain w' using a finite state transducer that consists of a phone error model, a lexicon and a language model. Alternatively, when untranscribed acoustic data are available, we can discover w' from the decoding results of a speech recognizer. The following two sections will discuss these two approaches in detail. In addition, a scaling constant α on the language model is used whose value depends on the dynamic ranges of $p(w)$ and $p(\phi|w)$.

3. INDUCTIVE APPROACH

In the inductive learning setting, our goal is to produce a discriminatively trained LM that is optimal for recognizing any utterance in the same domain. In contrast, a transductive learning setting (Section 4) only focuses on optimizing the test-set performance, which may or may not generalize to other in-domain data.

In this section, we approximate the sum over w in Equation (2) by the sum over all maximum-order n-grams in the initial LM $\tilde{p}(w)$:

$$J_1 = -\sum_{w\in\{\text{n-grams}\}}\sum_{\phi}\tilde{p}\left(w\right)\tilde{p}\left(\phi|w\right)\log\frac{p_\lambda(w)^\alpha p(\phi|w)}{\sum_{w'} p_\lambda(w')^\alpha p(\phi|w')}. \quad (4)$$

The lexicon $\tilde{p}(\phi|w)$ then generates the corresponding phone sequence ϕ for each maximum-order n-gram. The simplest way to generate a set of confusable word sequences w' is to find homophone sequences for w according to the lexicon. However, the resulting set would be very sparse (not many w have homephones). Instead, we adopt a phone error model to perturb the phone sequence, and map the perturbed phone sequences back to possible word sequences using a lexicon and a language model. A phone-to-word transducer is built to perform this process. Specifically, the transducer comprises three components:

1. E: a phone error model that encodes $p(\phi|\phi')$, where ϕ' is the corrupted phone sequence w.r.t. the intended one ϕ. In this work, we estimated a monophone error model from results of a phone recognition task using the same acoustic model;
2. P: a pronunciation lexicon, which encodes $\tilde{p}(\phi'|w)$;
3. L: a language model, which encodes $\tilde{p}(w)$.

Given an intended ϕ, the transduction process can be written as $w = \phi \circ E \circ P \circ L$. An output w comes with a path score $\sum_{\phi'} p(\phi|\phi')\tilde{p}(\phi'|w)\tilde{p}(w)$; and the top n outputs, ranked by their path scores, form a set of confusable word sequences. $p(\phi|w)$ inside the log in Equation (4) is computed by $\sum_{\phi'} p(\phi|\phi')\tilde{p}(\phi'|w)$.

3.1. Optimizing LM parameters λ

Having described all components in Equation (4), we now describe a stochastic gradient descent method of optimizing J_1. Denote the **log** probability of the k^{th} n-gram as λ_k. The gradient of (4) with respect to λ_k is given by

$$\frac{\partial J_1}{\partial \lambda_k} = -\sum_{w\in\{\text{n-grams}\}}\sum_{\phi}\tilde{p}\left(w\right)\tilde{p}\left(\phi|w\right)\cdot \left[C_k(w) - \sum_{w'} p(w'|\phi)C_k(w')\right], \quad (5)$$

where $C_k(w)$ is the number of times that the k^{th} n-gram occurs w. In each epoch of stochastic gradient descent, λ_k is added by an amount proportional to the gradient, *i.e.*, $\Delta_k = \eta(\partial J_1/\partial\lambda_k)$, where η is the step size. The parameters λ_k are iteratively updated until the change in the objective J_1 is smaller than a threshold. We update n-grams of the highest order. When backoff happens, however, we fix backoff weights and update lower-order n-grams. At each iteration, we do not normalize LM parameters, as normalization would require constrained optimization which is unnecessary to our task.

It is worth noting that we empirically found that the dynamic range of Δ_k can be rather large due to the large dynamic range of $\tilde{p}(w)$. To compensate for this problem, we use a normalized gradient by replacing $\tilde{p}\left(w\right)$ in (5) with $\tilde{p}\left(w\right)/\sum_{w'}\tilde{p}\left(w'\right)$. This normalization leads to a slightly better performance in practice compared to using the gradient without normalization.

3.2. Comparison to Kurata's approach

In [6], Kurata et al. proposed a discriminative training method of LMs without acoustic data. They generated pseudo-ASR n-best lists for a sampled set of training texts, and applied MCE training using the generated n-best lists. Our inductive approach is similar to theirs, but differs in the following aspects:

- Training data: Kurata et al. used a sampled set of training texts, while we view the maximum-order n-grams as "training sentences" to which the transducer is applied. In this way, we include all n-grams in discriminative training without transducing millions of text queries in our task.
- Transduction: Kurata et al. computed $p(\phi|\phi')$ based on acoustic model distance. We resort to the phone recognition statistics obtained directly from a phone recognizer. However, it is not clear to us which one gives a better representation of phonetic/acoustic confusability.
- Training objective: Kurata et al. applied MCE training while our approach is based on the minimum conditional entropy criterion.

4. TRANSDUCTIVE APPROACH

Transductive learning [8] is a machine learning paradigm that aims at minimizing the risk of the test set. In this setting, we assume

the availability of test-set inputs, denoted by $\{x_i\}_i^m$, and we desire to predict as accurately as possible their corresponding word sequences. The key characteristic of the transductive approach is that the resulting LM is optimized for the test set only, which may or may not generalize. In this regard, this setting essentially corresponds to an unsupervised adaptation scenario, where the test-set inputs are used to influence LM training.

Specifically, for each test-set utterance x_i, we use a speech recognizer to obtain a set of n-best word sequence hypotheses, denoted as R_i. Two word sequences are considered confusable with each other if they belong to the same n-best list. We let $\mathbf{R}$ represent the union of all R_i, $i = 1, 2, \ldots, m$. Next, we let ϕ_i denote the *true* phone sequence for x_i. We introduce a virtual lexicon $p^{\text{vir}}(\phi|w)$ in which each entry consists of a word sequence $w \in \mathbf{R}$ and its corresponding "pronunciation"; we consider ϕ_i as a pronunciation for w if and only if $w \in R_i$. Then, we replace the sum over all possible w in Equation (2) by the sum over $w \in \mathbf{R}$ only, and replace $\tilde{p}(\phi|w)$ by $p^{\text{vir}}(\phi|w)$. Consequently, the training objective becomes,

$$J_2 = -\sum_{w\in\mathbf{R}} \tilde{p}\,(w) \sum_{\phi} p^{\text{vir}}(\phi|w) \log p\,(w|\phi)\,. \tag{6}$$

Note that the concept of $p^{\text{vir}}(\phi|w)$ is introduced only for notional consistency. An equivalent representation of J_2 is given by

$$J_2 = -\sum_{w\in\mathbf{R}} \tilde{p}\,(w) \sum_{i=1}^{m} \frac{1(w \in R_i)}{\sum_{j=1}^{m} 1(w \in R_j)} \log p\,(w|\phi_i)\,, \tag{7}$$

where $1(\cdot)$ is an indicator function. Apparently, this objective requires the knowledge of ϕ_i which we do not have. In this work, we simply use $p(w|x_i)$ as a substitute, *i.e.*,

$$p(w|\phi_i) \leftarrow p(w|x_i) = \frac{p_\lambda(w)^\alpha p(x_i|w)}{\sum_{w'\in R_i} p_\lambda(w')^\alpha p(x_i|w')}, \tag{8}$$

where $p(x_i|w)$ is the acoustic model (AM) score; α is the language model scalar; and w' are obtained from the n-best list of each test-set utterance, representing a set of word sequences truly confusable to the recognizer. This is a key difference from the inductive approach where w' are fabricated using a transducer.

4.1. Optimizing LM parameters λ

Taking the derivative of J_w with respect to λ_k, we have

$$\frac{\partial J_2}{\partial \lambda_k} = -\sum_{i=1}^{m} \sum_{w\in R_i} \frac{\tilde{p}\,(w)}{\sum_{j=1}^{m} 1(w \in R_j)} d_k(w, i), \tag{9}$$

where $d_k(w,i) = C_k(w) - \sum_{w'\in R_i} p(w'|\phi_i) C_k(w')$. A delta, which equals the gradient scaled by a step size, is used to update λ_k in a *batch* mode — the gradient of λ_k is accumulated at each utterance, and is used to update λ_k after seeing all utterances.

Alternatively, we can update LM parameters in an *online* mode, meaning that we update λ at each utterance. Since only one utterance is considered at a time, we can eliminate $\sum_{j=1}^{m} 1(w \in R_j)$ in the denominator, leading to the following gradient at utterance i:

$$\frac{\partial J_{2,i}}{\partial \lambda_k} = -\sum_{w\in R_i} \tilde{p}(w) d_k(w, i). \tag{10}$$

In both cases, we perform the stochastic gradient descent iteratively until the objective value saturates.

4.2. Optimizing LM scalar α

In a similar way, we can update the LM scalar α. While this parameter is often tuned on a development set (α becomes a practical scalar in a recognizer when the AM score is used in Equation (8)), this work explores the possibility of optimizing α under the conditional entropy criterion on test-set inputs. The gradient of J_2 with respect to α, in a batch mode, is given by

$$\frac{\partial J_2}{\partial \alpha} = -\sum_{i=1}^{m} \sum_{w\in R_i} \frac{\tilde{p}\,(w)}{\sum_{j=1}^{m} 1(w \in R_j)} e\,(w, i)\,, \tag{11}$$

where $e(w,i) = \log p_\lambda(w) - \sum_{w'} p(w'|\phi_i) \log p_\lambda(w')$. An online version of Equation (11), similar to Equation (10), is obtained by eliminating $\sum_{j=1}^{m} 1(w \in R_j)$ in the denominator.

In fact, updating α alone can be viewed as tying all λ_k together and adjusting them all at once. This is computationally more efficient compared to updating a large number of λ_k. Additionally, we can update both α and λ_k, as will be shown in our experiments.

5. EXPERIMENTS

5.1. Test data

We evaluate our methods in a UVS task, *i.e.*, to recognize spoken queries in the general web search domain. Since this is a relatively new speech application, we do not have any transcribed in-domain acoustic data by the time of our experiments, to apply supervised discriminative training of language models. Although the language modeling techniques proposed in this work do not require transcribed acoustic data for training, we still need to collect a small set of such data for evaluating recognition performance. To this end, we randomly sampled text queries from the Bing search query log as prompts, and then asked 50 speakers to read the prompts over their mobile phones or land lines. The prompts are considered as the ground truth of the corresponding utterances. Our test set consists of 2075 utterances collected in this manner.

5.2. System setup

We deploy the same acoustic model as was used in [5] for voice-enabled local search. We use a combination of a hand-authored pronunciation lexicon and a letter-to-sound system to generate pronunciations for words. Our baseline LM is a trigram model with back-off, trained on millions of text queries issued to Bing search. After applying cutoffs at different levels, the final LM consists of 100K unigrams, 1.7M bigrams and 1.2M trigrams. In decoding, we use an initial LM scalar equal to 15, which was optimized for the voice search system in [5]. The decoder is a Gaussian-mixture HMM based speech recognizer implemented using HTK. The baseline recognizer gives a sentence accuracy of 67.71% on the test set.

Our proposed methods are evaluated on both rescoring and first-pass decoding experiments. For rescoring, the n-best lists output by the baseline recognizer are rescored using the newly-trained language model. An oracle experiment shows that the best rescoring accuracy that can be achieved using the n-best lists with n=10 is 80.00%. For first-pass decoding, we run the recognizer with the discriminatively-trained LM.

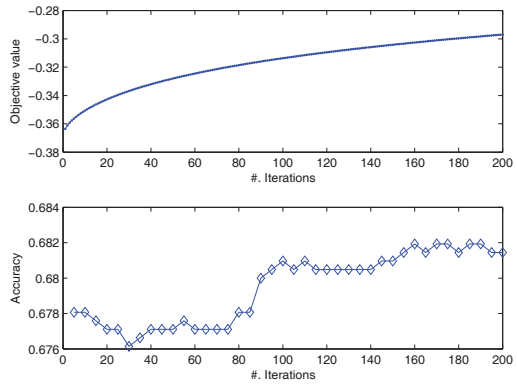

Fig. 1. Training objective value (upper panel) and recognition accuracy (lower panel) on the test set over iterations for the inductive approach.

Table 1. Sentence recognition accuracies (%) of the baseline LM and the discriminatively-trained LMs

	Rescoring	First-pass decoding
Baseline	67.71	67.71
Inductive approach		
Optimize λ	68.19	67.66
Transductive approach		
Optimize λ online	69.30	68.96
Optimize λ batch	68.53	68.33
Optimize α online	69.78	69.11
Optimize α batch	69.59	70.07
Optimize α then λ	70.17	70.31

5.3. Inductive approach results

For the inductive approach, the sum in Equation (4) was taken over all 1.2M trigrams in the LM, but any order of n-gram can be included with the same approach. We applied the stochastic gradient descent method described in Section 3 to update the baseline LM. We used an LM scalar $\alpha = 0.15$ and a step size $\eta = 10^{-2}$. The algorithm ran for 200 iterations before the objective value converged. Figure 1 shows the objective value and the sentence recognition accuracy for the rescoring results over iterations. Table 1 contains accuracies for both rescoring and first-pass decoding using the inductive approach. As shown, the inductive approach gives an absolute 0.48% accuracy increase in rescoring, while it does not help in first-pass decoding.

5.4. Transductive approach results

For the transductive approach, we first ran the baseline recognizer on the test-set utterances to generate n-best lists (n=10) and extracted both LM and AM scores for each utterance. Then the techniques described in Section 4 were used to update the baseline LM. We compared five settings for this approach. The first two settings correspond to optimizing λ only, one using the batch mode and the other using the online mode. The next two settings correspond to optimizing α only. In the last setting, we first update α in a batch mode, and then update λ in an online mode (other configurations gave comparable results). Table 1 summarizes the sentence accuracies for rescoring and first-pass decoding. Optimizing λ alone helps increase the accuracy by an absolute 0.62-1.59%. Optimizing α alone helps increase the accuracy by an absolute 1.40-2.36%. Optimizing α then λ gives the largest accuracy gain for both rescoring (absolute 2.46%) and first-pass (absolute 2.6%) decoding experiments.

One important factor that contributes to the effectiveness of the transductive approach is that the set of confusable word sequences are discovered by the recognizer directly. As a result, the LM discriminatively trained against such a set is optimized to overcome the confusability in real decoding. For the test set of 2075 utterances, one iteration took only several seconds using a single CPU.

6. CONCLUSION

This paper proposes discriminative training methods for language modeling without the need of transcribed audio data. Our approach aims at minimizing the conditional entropy of word sequences given phone sequences, which only requires the availability of an initial language model, a pronunciation lexicon, and a certain form of confusability information. This training criterion can be implemented in two settings. In an inductive learning setting, a phone-to-word transducer that incorporates a phone error model is used to generate confusable word sequences, while in a transductive learning setting, the decoded n-best lists of the test set form the confusable sentence set. The experiments showed a moderate accuracy improvement using the LM trained with the inductive approach, and significant improvements using the LM trained with the transductive approach in both rescoring and first-pass decoding.

For future work, we would like to apply our methods to other applications with longer utterances, to see how well they generalize. We are also interested in a a context-dependent phone error model for the inductive approach. The authors would like to thank Chin-Hui Lee for useful discussions, and thank Geoffrey Zweig for providing several tools used in the experiments.

7. REFERENCES

[1] H.-K. Jeff Kuo, Eric Fosle-Lussier, Hui Jiang, and Chin-Hui Lee, "Discriminative training of language models for speech recognition," in *Proc. ICASSP*, 2002, pp. 325–328.

[2] Jen-Wei Kuo and Berlin Chen, "Minimum word error based discriminative training of language models," in *Proc. INTERSPEECH*, 2005, pp. 1277–1280.

[3] Brian Roark, Murat Saraclar, and Michael Collins, "Discriminative n-gram language modeling," *Computer Speech and Language*, vol. 21, no. 2, pp. 373–392, 2007.

[4] Xiao Li, Patrick Nguyen, Geoffrey Zweig, and Dan Bohus, "Leveraging multiple query logs to improve language models for spoken query recognition," in *Proc. ICASSP*, 2009, pp. 3713–3716.

[5] A. Acero, N. Bernstein, R. Chambers, Y.C. Ju, X. Li, J. Odell, P. Nguyen, O. Scholz, and G. Zweig, "Live search for mobile:web services by voice on the cellphone," in *Proc. ICASSP*, 2008, pp. 5256–5259.

[6] Gakuto Kurata, Nobuyasu Itoh, and Masafumi Nishimura, "Acoustically discriminative training for language models," in *Proc. ICASSP*, 2009, pp. 4717–4720.

[7] Geoffrey Zweig and Jon Nedel, "Empirical properties of multilingual phone-to-word transduction," Tech. Rep. MSR-TR-2007-125, Microsoft Research, September 2007.

[8] V. Vapnik, *Statistical learning theory*, Wiley, 1998.