Lightly Supervised Training for Risk-Based Discriminative Language Models

Akio Kobayashi, Takahiro Oku, Yuya Fujita, Shoei Sato

NHK (Japan Broadcasting Corporation)
Science & Technology Research Laboratories, Tokyo, Japan
{kobayashi.a-fs, oku.t-le, fujita.y-gc, sato.s-gu}@nhk.or.jp

Abstract
We propose a lightly supervised training method for a discriminative language model (DLM) based on risk minimization criteria. In lightly supervised training, pseudo labels generated by automatic speech recognition (ASR) are used as references. However, as these labels usually include recognition errors, the discriminative models estimated from such faulty reference labels may degrade ASR performance. Therefore, an approach to prevent performance degradation is necessary for discriminative language modeling. In our proposed lightly supervised training, the DLM is estimated from a "fused" risk, which is a relaxed version of the conventional Bayes risk. The fused risk is computed in a supervised manner when pseudo labels are accepted as references with high confidence while computed in an unsupervised manner when the labels are rejected due to low confidence. Accordingly, minimizing the fused risk for the training lattices results in a DLM with smoothed model parameters. The experimental results show that our proposed lightly supervised training method significantly reduced the word error rate compared with DLMs trained in conventional lightly supervised manners.

Index Terms: discriminative training, language modeling, Bayes risk, lightly supervised training

1. Introduction

NHK (Japan Broadcasting Corp.) has developed a system for closed-captioning broadcast news using real-time automatic speech recognition (ASR) [1]. The system uses a hybrid approach involving a "direct method", which directly decodes original program sounds, and a "re-speak method", which decodes another speaker's rephrasing speech instead of the original sounds, especially in interviews and conversations. The re-speak method is also used solely for captioning live sports and information programs [2]. In the practical system, all the errors in the recognition results are fixed by operators manually in real-time. Thus, the availability of such applications strongly depends on the ASR performance. Recently, there has been much interest in applying discriminative acoustic and language models for achieving high performance. Although discriminative modeling typically requires a large amount of training data for obtaining a significant reduction in word error rate (WER), there are only limited resources available in reality mainly due to cost issues.

In ASR applications targeted at broadcast programs, however, transcriptions can be estimated with high accuracy using closed-captions associated with their program sounds. Such a task-dependent property leads to lightly supervised training, which estimates the statistical models using the estimated transcriptions as pseudo reference labels. Regarding acoustic modeling, many approaches have been proposed and achieved significant performance gains in ASR [3, 4, 5]. However, there has been less interest in discriminative language modeling than in acoustic modeling.

We propose a new lightly supervised training method for a discriminative language model (DLM) based on risk minimization in lightly supervised training for acoustic models, pseudo reference labels are generated using a biased language model (LM), which is adapted using a content-related corpus such as closed-captions [3, 4]. In discriminative language modeling, the model can be estimated similarly by using these pseudo labels as references in conventional supervised manners [6, 7]. However, as these labels usually contain recognition errors, the estimated DLM will most likely degrade ASR performance. Therefore, we extend conventional discriminative language modeling in the supervised/unsupervised manners to prevent performance degradation. The DLM parameters are estimated by minimizing an objective based on the expected risk similar to the Bayes risk [8]. Unlike previously proposed risk-based approaches [7, 9], we compute a "fused risk", which is an integration of risks obtained in supervised and unsupervised manners. The fused risk is computed in a supervised manner when the pseudo labels are regarded as references with high confidence while calculated in an unsupervised manner when the labels are rejected as "false" due to low confidence. Since a low confidence subset of pseudo labels is not accepted as references, the expected risk is partially computed using all the competing hypotheses and more relaxed than the risk obtained when whole labels are regarded as references. Consequently, minimizing the fused risk is expected to result in a DLM with smoothed model parameters.

We describe our novel method for discriminative language modeling using the fused risk in a lightly supervised manner and explain the experimental results from an evaluation of "re-speak" speech materials.

2. Discriminative Language Models

2.1. Overview

A discriminative language model (DLM) is described as a log-linear model [6, 10]. Given an audio input, \( x \), a posterior of sentence hypothesis, \( w \), is expressed as

\[
P(w|x; \Lambda) = \frac{1}{Z(\Lambda)} \exp \{ \lambda_{am} f_{am}(x|w) + \lambda_{lm} f_{lm}(w) + \sum_j \lambda_j f_j(w) \},
\]

where \( f_{am}(x|w) \) and \( f_{lm}(w) \) are scores given by acoustic/language models, respectively. In this paper, \( \lambda_{am} \) and \( \lambda_{lm} \) are constant weighting factors. The DLM is given by the last term, \( \sum \lambda_j f_j(w) \), where \( f_j \) denotes a feature function derived...
2.2. Risk-Based Discriminative Language Modeling

We now briefly review risk-based discriminative language modeling using labeled/unlabeled training data.

First, when we have training lattices with references (labeled data), an objective based on the Bayes risk [10] for DLM estimation is defined as

$$I_{\text{risk}}^{(l)}(\Lambda) = \frac{1}{M} \sum_{m=1}^{M} \sum_{w \in L_{m}} R(w_{m}^{\text{ref}}, w) P(w | x_{m}; \Lambda),$$  \hspace{1cm} (2)

where \(x_{m} \) \((m = 1, \ldots, M)\) is an audio input, \(R(w_{m}^{\text{ref}}, w)\) is a cost defined between the reference, \(w_{m}^{\text{ref}}\), and the hypothesis, \(w\), in the \(m\)-th training lattice, \(L_{m}\). Minimizing the objective with regard to \(\Lambda\) leads to a supervised-trained DLM.

If there are no references for \(x_{m}\) (unlabeled data), the above objective can be modified in an unsupervised manner [7].

$$I_{\text{risk}}^{(u)}(\Lambda) = \frac{1}{M} \sum_{m=1}^{M} \sum_{w \in L_{m}} P(w | x_{m}; \Lambda) \times \sum_{w \in L_{m}} R(w, w') P(w' | x_{m}; \Lambda).$$  \hspace{1cm} (3)

Note that Eq. (3) can be a generalized formula of Eq. (2) because all the hypotheses are regarded as references. The expected risks in Eqs. (2) and (3) are efficiently approximated on the training lattices by using edge-wise risks. In the supervised training setting, given a lattice, \(L\), and a reference, \(w_{\text{ref}}\), the edge-wise risk, \(\zeta^{(e)}(e)\), at the edge, \(e\), is expressed as

$$\zeta^{(e)}(e) = \sum_{e' \in w_{\text{ref}}} o(e, e') \ell_{0,1}(e, e').$$  \hspace{1cm} (4)

\(o(\cdot, \cdot)\) is an overlap function between two edges and defined as

$$o(e, e') = \min(\tau(e), \tau(e')) - \max(\sigma(e), \sigma(e')),$$  \hspace{1cm} (5)

where \(\sigma(e)\) denotes a start node for \(e\), while \(\tau(e)\) represents an end node. \(\ell_{0,1}(\cdot, \cdot)\) is a local cost function defined between overlapping edges. In this paper, \(\ell_{0,1}(\cdot, \cdot)\) is a simple binary function expressed as

$$\ell_{0,1}(e, e') = \begin{cases} 0 & \text{if label}(e) = \text{label}(e') \\ 1 & \text{otherwise} \end{cases}.$$  \hspace{1cm} (6)

Similarly, for the unlabeled training lattices, the edge-wise risk is expressed as

$$\zeta^{(u)}(e) = \sum_{e' \in L} o(e, e') \ell_{0,1}(e, e') p(e'),$$  \hspace{1cm} (7)

where \(p(e')\) is an edge posterior obtained using the forward-backward algorithm.

Accumulating the edge-wise risks by using the forward-backward algorithm, we obtain the entire expected risk of the lattice (Eqs. (2) and (3)) in a similar manner to minimum phone error training [11].

The model parameters, \(\Lambda\), are estimated by using quasi-Newton methods such as the L-BFGS algorithm [12]. According to [7], the gradient of the expected risk, \(\Delta_{\psi}\), which is required for the L-BFGS algorithm, is approximated by

$$\Delta_{\psi} = \frac{1}{M} \sum_{m=1}^{M} \sum_{e \in L_{m}} p(\psi) (\Upsilon_{m} - v(e)) \varphi_{\psi}(e),$$  \hspace{1cm} (8)

where \(\Upsilon_{m}\) denotes an expected risk of the \(m\)-th training lattice, \(v(e)\) is the expected risk of all the paths passing through \(e\), and \(\varphi_{\psi}(e)\) is a binary function that returns 1 if \(f_{\psi}\) is activated on \(e\). A detailed derivation procedure can be found in [7].

2.3. Lightly Supervised Training for DLM

In lightly supervised training, the recognition results (pseudo labels) are provided as the references for statistical modeling. Under this condition, the following strategies can be taken to estimate the DLMs.

St-I) Supervised training when all the pseudo labels are regarded as the references. In this strategy, model parameters, \(\Lambda\), are estimated using the objective Eq. (2). The edge-wise risks are computed based on Eq. (4).

St-II) Unsupervised training using a fused risk (see section 2.3.1). The edge-wise risks are computed based on Eqs. (4) and (7). This is regarded as a relaxed strategy of supervised training St-I.

2.3.1. Fused Risk

The fused risk is defined as an integration of the edge-wise risks expressed in Eqs. (4) and (7) so that it relaxes the risk obtained from only the labeled data.

$$\zeta^{(p)}(e) = \sum_{e' \in L_{m}, \psi(e') > \alpha} o(e, e') \ell_{0,1}(e, e') + \sum_{e'' \in L} o(e, e'') \ell_{0,1}(e, e'') p(e''),$$  \hspace{1cm} (9)

where \(\psi^{p}\) denotes a pseudo label sequence and \(c(\cdot)\) gives a confidence score of \(e'\). We used edge posteriors derived from the forward-backward algorithm as confidence scores. The first term is a modified version of Eq. (4) in that edge-wise risks are computed over the qualified reference edges having higher confidence scores than a threshold \(\alpha\). The second term is derived from Eq. (7), where \(o(\cdot, \cdot)\) is a modified overlap function, which returns an overlapping frame ratio excluding frames overlapped with qualified reference edges.

Figure 1 shows an example of fused risk computation. The currently focused edge, \(e\), is overlapped by \(m/n\) for the corresponding reference edge, \(e_{1}\). The partial edge-wise risk is
expressed as \( \zeta(p)(e) = o(e, e) f_{0,1}(e, e_1) \) because there is an overlapped qualified reference edge. As \( e_2 \) is not a reference because of low confidence, the other partial edge-wise risk is expressed as \( \zeta_2^{(w)}(e) = o(e, e_2) f_{0,1}(e, e_2) p(e_2) + o(e, e_3) f_{0,1}(e, e_3) p(e_3) \). In the case, \( o(e, e_2) \) returns \( k/n \) instead of \( (m+k)/n \) (Fig. 1). Finally, the fused risk is computed as \( \zeta(p)(e) = \zeta(p)(e) + \zeta_2^{(w)}(e) \).

2.3.2. Lightly Supervised Training

According to the following procedure, the DLM is obtained in a lightly supervised manner (SI-II). First, pseudo labels are generated using a biased LM, which is obtained by count-merge LM adaptation using closed-captions [13]. Next, confidence scores (edge posteriors) of word hypotheses on decoded lattices are calculated. The pseudo labels with confidence scores are tagged as “true” references when their confidence scores are higher than a pre-defined threshold and “false” otherwise. The approximate risk of the lattice is computed using the forward-backward algorithm by accumulating the edge-wise fused risks. The solution of the minimization problem in Eq. (3) leads to a set of DLM parameters, \( \Lambda \).

2.4. Conventional Discriminative Language Modeling

This section introduces two conventional approaches for discriminative language modeling in the supervised manner. These approaches are compared with the quasi-Newton algorithms for 

2.4.1. Conditional Log-Likelihood Maximization

One of the conventional discriminative language modeling used for labeled lattices is based on minimization of the negative conditional log-likelihood (CLL) [6]. The objective for parameter estimation is defined as

\[
\text{CLL}^{(1)}(\Lambda) = - \frac{1}{M} \sum_{m=1}^{M} \log P(w_m'|x_m; \Lambda). \tag{10}
\]

This objective is analogous to that used in maximum mutual information acoustic modeling. The model parameters are estimated using the quasi-Newton algorithms.

2.4.2. Large Margin Training

The other approach is based on large margin training. In large margin training, \( \Lambda \) are estimated by maximizing the following margin [14, 15].

\[
\forall w^\text{ref} \neq w, \quad D(x, w^\text{ref}; \Lambda) - D(x, w; \Lambda) + \rho R(w^\text{ref}, w), \tag{11}
\]

where \( D(x, w; \Lambda) \) is a score or an exponential part of Eq. (1),

\[
D(x, w; \Lambda) = \lambda_{\text{sm}} f_{\text{sm}}(x|w) + \lambda_{\text{lm}} f_{\text{lm}}(w) + \sum_j \lambda_j f_j(w), \tag{12}
\]

and \( \rho \) is a positive factor.

The model parameters are estimated using the perceptron-based algorithm [16]. First, the best hypothesis, \( w^b \), is obtained by Viterbi decoding (lattice rescoring) using the current parameters. Then, the gradient-based update formula is applied.

\[
\lambda_j \leftarrow \lambda_j + \kappa \frac{\partial}{\partial \lambda_j} \left[ D(x, w^\text{ref}; \Lambda) - D(x, w^b; \Lambda) \right], \tag{13}
\]

where \( \kappa \) is a learning rate.

3. Experiments

3.1. Setup

NHK’s speech decoder transcribes input audio streams in real time while detecting start and end points of speech segments [17]. The acoustic inputs are parameterized into 39 dimensional vectors: 12 mel frequency cepstral coefficients (MFCCs) with log-power and their first- and second-order differentials. The decoder uses a two-pass strategy that obtains 500-best sentence hypotheses by using gender-dependent tree lexica and a bigram LM in the first pass and rescors them using a trigram LM and a DLM. The acoustic model was trained from Japanese broadcast news, which consisted of 250 hours of female utterances. The speaker-adapted models were then obtained by maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) [18]. The baseline trigram LM was trained on Japanese broadcast news manuscripts, transcriptions and closed-captions (412M words), and the vocabulary size was set to 100k.

Table 1 lists the evaluation data, taken from female re-speaker’s speech for 11 episodes of an NHK information program. Six episodes were used as test data, and the remaining five episodes were used as development data. The perplexities (PP), out of vocabulary (OOV) rates, and WERs were measured using the baseline trigram LM.

Table 2 lists the training data for discriminative language modeling. The data were taken from re-speaker’s speech consisting of 105 episodes from the same information program. Discriminative language modeling was conducted on the decoded lattices of the training data. The feature functions were defined by word tuples (bigrams and trigrams) observed more than five times in the training lattices (Table 3). The biased LMs for generating the pseudo labels were constructed from trigram counts of the previously noted baseline corpus and the closed-captions (78.0k sentences, 110.6k words) that contain the extra captions for offline video footages unrelated to re-speaker’s speech. In the experiments, the biased LMs were estimated by count-merge LM adaptation.

3.2. Experimental Results

Table 4 lists the WER results for the evaluation data. In the table, we compared the following DLMs.

**Baseline:** The baseline trigram LM (w/o DLMs).

**Risk:** The DLM trained by risk minimization.

**CLL:** The DLM trained by negative conditional log-likelihood minimization.

**Margin:** The DLM based on large margin training.
In lightly supervised training (St-II), the thresholds for reference selection, $\alpha$, were varied among the values from 0.00 to 1.00 (denoted as values in parentheses). When $\alpha$ is set to 0.00, the DLM is estimated by regarding all the pseudo labels as "true" references; thus, it is equivalent to supervised training (St-I). When $\alpha$ is set to 1.00, the DLM is trained by rejecting all the labels as "false" (equivalent to unsupervised training).

For the test data, Risk (supervised) achieved a WER of 11.2% and a relative reduction of 3.4% compared with Baseline. It was a significant improvement according to a matched-pairs test [19] and the best performance among all DLMs. CLL gave a similar significant reduction in supervised training while Margin did not reach the best performance because the perceptron-based algorithm does not estimate all the model parameters in principle. Without any references in the training data, the risk-based DLM (unsupervised) exhibited less improvement on the test data, which was a WER of 11.5%. Compared with the result from Risk (supervised), there is an absolute gap of 0.3% (from 11.5% to 11.3%) to be recovered by the lightly supervised trained DLM.

In St-I, where all the pseudo labels are accepted as references, Risk and CLL achieved WERs of 11.4%. On the other hand, in St-II, Fused DLMs achieved WERs of 11.3% when thresholds were set to 0.99 and 0.90 and recovered 0.2% absolute compared with Risk (unsupervised). In these settings, 77.1% of all the pseudo labels were accepted as "true" references where $\alpha = 0.99$ while 88.3% were used where $\alpha = 0.90$. Although the result from Fused (0.99) exhibited a small WER reduction (0.1% absolute) compared with Risk (St-I), a matched-pair test showed that the WER was decreased at a significance level of 0.05. Therefore, our proposed training method was proved to be effective for reducing WER.

### Table 3: Feature Functions

<table>
<thead>
<tr>
<th></th>
<th>bigrams</th>
<th>trigrams</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>181.6k</td>
<td>219.4k</td>
<td>401.0k</td>
</tr>
<tr>
<td>unsupervised</td>
<td>181.6k</td>
<td>217.0k</td>
<td>399.4k</td>
</tr>
<tr>
<td>lightly supervised</td>
<td>182.8k</td>
<td>221.0k</td>
<td>403.8k</td>
</tr>
</tbody>
</table>

### Table 4: Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>Dev.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>12.3</td>
<td>11.6</td>
</tr>
<tr>
<td>Labeled (supervised)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>11.8</td>
<td>11.2</td>
</tr>
<tr>
<td>CLL</td>
<td>12.0</td>
<td>11.3</td>
</tr>
<tr>
<td>Margin</td>
<td>12.1</td>
<td>11.4</td>
</tr>
<tr>
<td>Unlabeled (unsupervised)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>12.1</td>
<td>11.5</td>
</tr>
<tr>
<td>St-I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>11.9</td>
<td>11.4</td>
</tr>
<tr>
<td>CLL</td>
<td>11.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Margin</td>
<td>12.1</td>
<td>11.5</td>
</tr>
<tr>
<td>Pseudo-Labeled (lighty supervised)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>St-II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>11.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Fused (0.00)</td>
<td>11.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Fused (0.30)</td>
<td>11.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Fused (0.70)</td>
<td>11.8</td>
<td>11.4</td>
</tr>
<tr>
<td>Fused (0.90)</td>
<td>11.8</td>
<td>11.3</td>
</tr>
<tr>
<td>Fused (0.99)</td>
<td>11.9</td>
<td>11.3</td>
</tr>
<tr>
<td>Fused (1.00)</td>
<td>12.1</td>
<td>11.5</td>
</tr>
</tbody>
</table>

### Table 5: Detailed Results for Test Data

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>SUB</th>
<th>DEL</th>
<th>INS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>11.6</td>
<td>7.9</td>
<td>1.5</td>
<td>2.2</td>
</tr>
<tr>
<td>Risk (supervised)</td>
<td>11.2</td>
<td>7.7</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Risk (unsupervised)</td>
<td>11.5</td>
<td>8.0</td>
<td>1.7</td>
<td>1.9</td>
</tr>
<tr>
<td>Fused (0.00, St-I)</td>
<td>11.4</td>
<td>7.8</td>
<td>1.5</td>
<td>2.1</td>
</tr>
<tr>
<td>Fused (0.99, St-II)</td>
<td>11.3</td>
<td>7.8</td>
<td>1.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

### 4. Discussion

The overall results from the Fused DLMs shown in Table 4 indicate that model parameter estimation based on the fused risks can contribute to improving ASR performance. Then, we investigated the detailed results including substitution, deletion and insertion error rates for the test data in order to confirm the efficacy of fused risk (Table 5). In the table, the risk-based DLM (unsupervised) achieved the smallest insertion error rate of 1.9%. As discussed in [20], the risk-based DLM trained in the unsupervised manner tends to reduce the insertion errors while it increases the deletion errors. This tendency typically arises from the risks for word hypotheses consisting of a small number of phonemes, i.e., short words such as particles in Japanese. Since these words could be easily confused with similar words, they are expected to have larger risks than others. As a result, the estimated model parameters are likely to be negative to penalize the short words as insertions.

The same goes for our proposed lightly supervised training method since the fused risk is computed partially in the unsupervised manner by definition. When the results of Fused (0.00) and Fused (0.99) in Table 5 were compared, the insertion error rate was reduced by 0.1% absolute. It suggests that the DLM estimated from the fused risk would reflect the tendency observed in the unsupervised risk-based DLM. In lightly supervised training, the risks for short words are expected to be obtained in the unsupervised manner because corresponding pseudo labels are not accepted as references due to their low confidence scores. The entire risk of the lattice is naturally more relaxed than the risk obtained where all the pseudo labels are accepted. Consequently, minimizing the fused risk would lead to a smoothed DLM, which enables the reduction of insertion errors. However, the experimental results were from the DLMs trained using a comparatively small amount of training data; thus a further experimental study should be conducted when a larger amount of training data and more efficient confidence measures are used.

### 5. Conclusions

We proposed a lightly supervised training method for DLMs. Our proposed method was designed to improve ASR performance using a “fused” risk-based DLM. The fused risk is derived from the integration of risks obtained in supervised/unsupervised manners, and minimization of the risk results in a smoothed DLM. Experimental results showed that our lightly supervised trained DLMs achieved promising results compared with those trained in the conventional lightly supervised manner. The proposed method could be easily applied to discriminative acoustic modeling based on the Bayes risk. We will examine acoustic modeling based on the proposed lightly supervised training method in future work.
6. References


