A COMPARATIVE STUDY ON METHODS OF WEIGHTED LANGUAGE MODEL TRAINING FOR RERANKING LVCSR N-BEST HYPOTHESES

Takanobu Oba, Takaaki Hori and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation.
{oba, hori, ats}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper focuses on discriminative n-gram language models for a large vocabulary speech recognition task. Specifically we compare three training methods, Reranking Boosting (ReBst), Minimum Error Rate Training (MERT) and the Weighted Global Log-Linear Model (W-GCLM). They have a mechanism for handling sample weights, which are useful for providing an accurate model and work as impact factors of hypotheses for training. W-GCLM is proposed in this paper. We discuss the relationship between the three methods by comparing their loss functions. We also compare them experimentally by reranking N-best hypotheses under several conditions. We show that MERT and W-GCLM are different types of expansion of ReBst and have different respective advantages. Our experimental results reveal that W-GCLM outperforms ReBst and whether MERT or W-GCLM is superior depends on the training and test conditions.

Index Terms— Weighted GCLM, Reranking Boost, MERT, Discriminative LM, Error Correction

1. INTRODUCTION

In general, a speech recognition result is a word sequence with the maximum score among many hypotheses generated by a speech recognizer. However, these hypotheses, in short a word N-best or a lattice, include some word sequences with lower word error rates (WER) than the recognition result. One of the most typical ways to obtain such lower WER word sequences is the reranking/rescoring approach. Reranking is done by giving each hypothesis an extra score.

We focus on discriminative language models (DLM) based on a log-linear model [1, 2, 3, 4, 5, 6, 7, 8]. The models include information about pairs of reference and error words as a result of training using recognition hypotheses and is applied to reranking. So DLM based reranking is often called error correction approach. The parameter estimation is formalized as a minimization problem of a predefined loss function. Loss functions are usually constructed based on a margin, which is the distance between a hypothesis and its reference. That is, the purpose of training is to give large scores to word sequences with low WERs, and to give small scores to the others. Word n-gram is widely used as a feature.

In this paper, we compare some training methods where each hypothesis has a sample weight. Sample weights are indicators of important hypotheses for parameter estimation. Therefore, by employing proper sample weights, we can obtain a more accurate model than by training without sample weights. Typically WERs are used for this purpose. Some training methods with sample weights have been proposed in the machine learning area. One of them has been applied recently to error correction for speech recognition [8]. However, neither the relationship between the speech recognition training methods nor their performance has been sufficiently discussed.

First we describe other research on n-gram based error correction to clarify the relationship with our work described in this paper. In 2004, error correction was employed for speech recognition by Roark, et al., [1, 2, 6]. They used the Global Conditional Log-linear Model (GCLM) for parameter estimation. In 2006, Zhou, et al., compared some of the most typical reranking algorithms [4]. They reported that reranking boosting (ReBst) provided the most accurate model. In the above studies, no sample weights are considered in model training. In 2008, Kobayashi, et al., introduced a minimum WER training method for speech recognition [8]. This is an example of training where sample weights are given. Our work corresponds to Zhou’s work except that sample weights are given in our approach.

In this paper, we compare three methods, ReBst, the minimum error rate training (MERT) and Weighted GCLM (W-GCLM). ReBst was proposed by Collins, et al., [9]. ReBst has an architecture for handling sample weights, although sample weights were not given in the Zhou’s work. MERT was proposed by Och [10] and was used for speech recognition by Kobayashi. W-GCLM is an expansion of GCLM and is proposed in this paper.

We first discuss the three methods through a comparison of their loss functions. We indicate that; MERT and W-GCLM are different types of expansion of ReBst, W-GCLM can outperform ReBst in principle, MERT and W-GCLM include different mechanisms for constructing an accurate model.
We next compare the three methods experimentally. We applied each method to training for the reranking of 5000-best hypotheses under several conditions using the Corpus of Spontaneous Japanese (CSJ) [11]. W-GCLM outperformed ReBst under all the conditions. Our comparison of MERT and W-GCLM showed that the better performance depends on the relationship between training and test data. We also compared two types of sample weights, WERs and ranks. The use of the rank weights realized a model that was at least as accurate as a model generated using the WER weights.

This paper is organized as follows: in section 2, we describe the basics of error correction and the two conventional training methods, ReBst and MERT. W-GCLM is proposed in section 3. In section 4, we discuss the relationship between the three methods. Section 5 provides our experimental results. And section 6 concludes this paper.

2. RERANKING HYPOTHESES AND CONVENTIONAL TRAINING METHODS

2.1. Reranking Hypotheses

We represent N-best hypotheses generated from a speech recognition system as \( L = \{h_j|j = 1, 2, \cdots, N\} \), and a feature vector of a hypothesis \( h \) as \( f(h) \). Specifically, the speech recognition score of \( h \) is denoted as \( f_0(h) \).

Where \( a \) is a given parameter vector, the goal of the reranking problem is to find a hypothesis with the highest score. That is formulated as follows.

\[
h^* = \arg \max_{h \in L} \{a_0 f_0(h) + a^\top f(h)\}
\]

where \( a_0 \) is a given scaling constant. \(^\top\) denotes the transpose of the matrix.

2.2. Conventional Reranking Model Training Methods

For training, we prepare a data set that comprises;

- N-best lists \( \{L_i|i = 1, 2, \cdots, I\} \)
  - Each hypothesis is converted to a feature vector, which is denoted as \( f_{i,j} \). That is, \( L_i = \{f_{i,j}|j = 1, 2, \cdots, N_i\} \).
  - Sample weights
    - Each hypothesis has a sample weight \( w_{i,j} \).
  - References
    - We represent the feature vector of a reference as \( f_{i,0} \).

The parameters are estimated by finding a parameter vector \( a \) that minimizes a predefined loss function. Loss functions are usually constructed based on the margin

\[
M_{i,j} = a^\top f_{i,0} - a^\top f_{i,j}.
\]

[ReRanking Boost with Exponential Loss]

The loss function of ReBst is as below.

\[
L_{\text{ReBst}} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} w_{i,j} \exp(-M_{i,j})
\]

An efficient algorithm has been proposed for estimating \( a \) in a case with Boolean features [9].

[Minimum Error Rate Training]

The loss function of MERT is as follows.

\[
L_{\text{MERT}} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} w_{i,j} \exp(-M_{i,j})^\alpha
\]

When \( w_{i,j} \) is the error rate, the loss function corresponds to the expectation value of the error rate. \( \alpha \) is a hyper-parameter to balance the number of local optimal solutions with the potential minimum value of the loss function.

3. WEIGHTED GLOBAL CONDITIONAL LOG-LINEAR MODELS

We propose W-GCLM, which is an expanded GCLM for weighted training. The loss function is defined as follows.

\[
L_{\text{WGCLM}} = -\sum_{i=1}^{I} \log \frac{\exp(a^\top f_{i,0})}{\sum_j \exp(a^\top f_{i,j})}
\]

\[
= -\sum_{i=1}^{I} \log \left( \sum_j w_{i,j} \exp(-M_{i,j}) \right)
\]

This function is concave, which follows from the concavity of the functions of the form \( F(x) = \sum_i \log \sum_j \exp(x^\top c_{i,j}) \). Since \( w_{i,j} \exp(-M_{i,j}) = \exp(a^\top \{f_{i,j} - f_{i,0}\} + a_K \log w_{i,j}) \) where \( a_K = 1 \), if \( a_K \) is a variable, this function is explicitly concave. Even if \( a_K \) is constant the concavity for \( a \) is ensured. Therefore \( L_{\text{WGCLM}} \) is concave for \( a \).

For parameter estimation, in practice prior distributions should be considered to avoid the risk of overfitting for a training set. That is,

\[
L_{\text{WGCLM}} = -\sum_{i=1}^{I} \log \frac{\exp(a^\top f_{i,0})}{\sum_j \exp(a^\top f_{i,j})} + \frac{\|a\|}{2C}
\]

\( C \) is the regularization constant. We use L-2 norm.

4. RELATIONSHIP AMONG LOSS FUNCTIONS

When making connections between the loss functions of ReBst, MERT and W-GCLM, it is useful to consider the error count loss function,

\[
L_{\text{EC}} = \sum_{i=1}^{I} \sum_j \left[ -M_{i,j} \right]
\]
where $\lfloor x \rfloor = 0$ if $x$ is negative, 1 otherwise.

For weighted training, the error count loss can be expanded in the following two ways.

$$
L_{\text{WEC1}} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} w_{i,j} [-M_{i,j}] 
$$

(9)

$$
L_{\text{WEC2}} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} [-M_{i,j} + w_{i,j}] 
$$

(10)

In the latter case, the sample weights play a margin bias role. Since each weight directly affects the margin, it is expected that the latter loss is more effective for making an accurate model than the former loss.

The loss of ReBst is the exponential form of the weighted error count losses. Equation (3) can easily be constructed by replacing $\lfloor x \rfloor$ in equation (9) with $\exp(x)$. Furthermore, by rewriting equation (3) as

$$
L_{\text{ReBst}} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} \exp(-M_{i,j} + \log w_{i,j}), 
$$

(11)

we can understand that the weights in ReBst also play a margin bias role as in equation (10).

The loss of MERT corresponds to the loss of ReBst with the normalization term. This normalization leads to training without the reference feature vectors since $\exp(-a^T f_{i,0})$ is cancelled in the numerator and denominator, that is, equation (4) can be rewritten as

$$
L_{\text{MERT}} = \sum_{i=1}^{I} \sum_{j=1}^{N_i} w_{i,j} \frac{\exp(a^T f_{i,j})^\alpha}{\sum_{j=1}^{N_i} \exp(a^T f_{i,j})^\alpha}. 
$$

(12)

It is also important to mention that losing the reference vectors means losing the margins, in short, the weights lose their margin bias role.

The loss of W-GCLM (equation (6)) is constructed by introducing the function $\log(x)$ to the loss of ReBst (equation 3). The introduction of $\log(x)$ leads to the concavity of the loss function and normalizes W-GCLM. Equations (5) and (12) show the difference between the normalization methods of W-GCLM and MERT. In addition, in contrast to MERT the weights in W-GCLM contain a margin bias. That is, equation (6) can be rewritten as

$$
L_{\text{WGCLM}} = \sum_{i=1}^{I} \log \left\{ \sum_{j=1}^{N_i} \exp(-M_{i,j} + \log w_{i,j}) \right\}. 
$$

(13)

Table 1 summarizes the characteristics of the three methods. The methods all have mechanisms for generating an accurate model. ReBst and W-GCLM use sample weights as the margin bias. W-GCLM also has the merit of convergence to the global optimum. Therefore, W-GCLM outperforms ReBst in principle. On the other hand, MERT and W-GCLM have quit different mechanisms, although the both methods are expansions of ReBst. So it is difficult to say which is better. However, it is expected that W-GCLM generates a model that is more accurate but more sensitive to differences of tasks than MERT, because W-GCLM, which needs reference feature vectors for training, estimates the parameter vector so as to give large scores only to the references in the training set.

5. EXPERIMENTS

We used CSJ for our experiments. CSJ includes many lectures and their transcriptions. The lectures consist of academic and simulated presentations.

Table 2 shows the amount of data in our experimental environment. To make 5000-best lists, the utterances were recognized by using the speech recognition system SOLON, which was developed at NTT CSLabs. SOLON is a decoder based on a weighted finite state transducer and it can provide a fast efficient search by using a fast on-the-fly composition algorithm [12]. The acoustic model consists of MCE trained tri-phone HMMs with 5,000 states and 32 Gaussians, and the language model is a tri-gram model with Kneser-Ney smoothing.

Test set 2 and its development set consist of simulated lectures while the others consist of academic lectures. Test-set perplexities were calculated by using the language model of the SOLON. The perplexities of the training set and test set 1 are very close. We can expect test set 1 to include very similar linguistic features to the training set, although the perplexity does not explicitly express the closeness. At least we know that test set 2 includes very different features, since its perplexity is much bigger than the others.

Each development set is used to decide the scaling constant $a_0$ in equation (1), $\alpha$ in MERT, the regularization constant $C$ and the convergence conditions. The model selected using the development set is applied for the corresponding test set.

The WERs of the 1-best result of test sets 1 and 2 are 18.0% and 34.5%, respectively. Table 3 shows the results after applying reranking models. We used word uni-, bi-, and tri-gram Boolean features.

We first compare the models trained without sample weights. The results are shown in $w_{i,j} = 1$ column. MERT wasexcepted because its training always required sample

| Table 1. Relationship of ReBst, MERT and W-GCLM. |
|-----------------|-----------------|-----------------|
| training with reference feature vectors | ReBst | MERT | W-GCLM |
| margin bias | √ | - | √ |
| loss = evaluation function | - | √ | - |
| global optimum | - | - | √ |
weights. W-GCLM corresponds to plain GCLM in this situation. The result shows that (W-)GCLM outperforms ReBst. This is an effect of the convergence to the global optimum.

Next we compare the models trained where the sample weights are WERs. The results are shown in the “WER” column. The accuracies of ReBst and MERT are very close in test set 1, which is a linguistically-similar task to training set. The merits of ReBst and MERT described in table 1 provided almost the same increase in accuracy under this condition. The most accurate error correction was achieved by using W-GCLM.

W-GCLM also outperformed ReBst in test set 2 just as in test set 1. However, MERT generated the most accurate model for test set 2. This result shows that, when we compare MERT and W-GCLM, the more suitable method depends on the distance between the training and test sets. This result suggests that W-GCLM provides a model that is more sensitive to differences of tasks than MERT.

We next compare the models trained using ranks instead of WERs as weighting samples. The ranks are decided in order of WER in each N-best list. This is an effect of the convergence to the global optimum. W-GCLM outperformed ReBst both in principal and experimentally. And we found that the superiority of MERT versus W-GCLM depended on the training and test environment.

Table 2. Experimental data. Dev. lect., uttr. and p.p. denote development, lectures, utterances and perplexity, respectively.

<table>
<thead>
<tr>
<th>training set</th>
<th>lecture</th>
<th># of lect. (attr.)</th>
<th># of words</th>
<th>p.p.</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev. set 1</td>
<td>academic</td>
<td>150 (25, 130)</td>
<td>420,918</td>
<td>68.7</td>
</tr>
<tr>
<td>test set 1</td>
<td>academic</td>
<td>10 (1, 293)</td>
<td>26,329</td>
<td>76.1</td>
</tr>
<tr>
<td>dev. set 2</td>
<td>simulated</td>
<td>10 (1, 479)</td>
<td>20,990</td>
<td>96.1</td>
</tr>
<tr>
<td>test set 2</td>
<td>simulated</td>
<td>10 (717)</td>
<td>17,242</td>
<td>142.3</td>
</tr>
</tbody>
</table>

Table 3. WER

<table>
<thead>
<tr>
<th>sample weight (w_{i,j})</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReBst</td>
<td>17.9</td>
<td>34.0</td>
</tr>
<tr>
<td>MERT</td>
<td>17.7</td>
<td>33.1</td>
</tr>
<tr>
<td>W-GCLM</td>
<td>17.6</td>
<td>32.9</td>
</tr>
</tbody>
</table>

6. CONCLUSION

We focused on the error correction of speech recognition and compared three training methods, ReBst, MERT and W-GCLM, where sample weights are given. W-GCLM is a expansion of GCLM and is proposed in this paper. We also discussed the relationship between the three methods by comparing their loss functions. We indicated that W-GCLM outperformed ReBst both in principal and experimentally. And we found that the superiority of MERT versus W-GCLM depended on the training and test environment.

7. REFERENCES