MULTI-OBJECTIVE OPTIMIZATION FOR SEMI-SUPERVISED DISCRIMINATIVE LANGUAGE MODELING

Akio Kobayashi, Takahiro Oku, Toru Imai

NHK Science and Technology
Research Laboratories
Tokyo, Japan.

Seiichi Nakagawa

Toyohashi University of Technology
Toyohashi, Japan.

ABSTRACT

A method for semi-supervised language modeling, which was designed to improve the robustness of a language model (LM) obtained from manually transcribed (labeled) data, is proposed. The LM is implemented as a log-linear model, which employs a set of linguistic features derived from word or phoneme n-grams. The proposed method is formulated as a multi-objective optimization programming problem (MOP), which consists of two objective functions based on expected risks for labeled lattices and automatic speech recognition (ASR) lattices as unlabeled training data. The model is trained in a discriminative manner and acquired as a solution to the problem. In transcribing Japanese broadcast programs, the proposed method reduced word error rate by 6.3% compared with that achieved by a conventional trigram LM.

Index Terms— discriminative training, semi-supervised training, language modeling, Bayes risk minimization

1. INTRODUCTION

The recent progress in the field of corpus-based spoken-language processing has led to its successful application in the real world. NHK (Japan Broadcasting Corp.) has developed a system for closed-captioning broadcast news using real-time automatic speech recognition (ASR) [1]. ASR technology also plays an important role in the development of a broadcast archiving system, which serves as a basis for spoken document processing applications. The availability of these applications strongly depends on the accuracy of ASR, and recently there has been many interest in applying discriminative acoustic or language models for improvement. Although these models typically require a large amount of manually transcribed (labeled) data, there are only limited resources available in reality. Information from unlabeled data such as ASR transcriptions could therefore be useful for increasing the robustness of the models.

In regard to acoustic modeling, many semi-supervised approaches for dealing with unlabeled data have been proposed [2, 3]. In most previous works, the acoustic models are obtained from labeled data in a discriminative manner, while unlabeled data are incorporated in the models through an objective function, which is defined in a generative manner. In the field of language modeling, such semi-supervised training has not been well studied so far, though supervised discriminative modeling [4, 5] and unsupervised modeling [6, 7] have been investigated individually.

Under these circumstances, our previous work on unsupervised LM adaptation with a single objective function [8] is expanded to deal with multiple objectives, and a novel semi-supervised language modeling method is proposed. The proposed method is formulated as a multi-objective optimization programming (MOP) problem, which reflects contributions from labeled and unlabeled training data to a language model flexibly. The problem employs objective functions that are designed for minimizing expected risks associated with word error rate (WER), and its optimum solution leads to a discriminatively-trained model. The proposed semi-supervised language modeling method is discussed and tested in an evaluation of Japanese broadcast transcriptions.

2. SEMI-SUPERVISED LANGUAGE MODELING

2.1. Log-Linear Language Models

A log-linear formulation of a language model (LM) is introduced in the following.

In ASR, on the basis of the Bayes' rule, the optimum sentence hypothesis, $\hat{\boldsymbol{w}}$, is given by

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x}) = \arg\max_{\boldsymbol{w}} P(\boldsymbol{x}|\boldsymbol{w})P(\boldsymbol{w}), \quad (1)$$

where $P(\boldsymbol{w}|\boldsymbol{x})$ is a posterior of sentence hypothesis, $\boldsymbol{w}$, given an audio input, $\boldsymbol{x}$. $P(\boldsymbol{x}|\boldsymbol{w})$ and $P(\boldsymbol{w})$ are given by an acoustic model and a language model, respectively. To reflect information from labeled and unlabeled training lattices, the posterior probability is redefined by using a set of linguistic features as

$$P(\boldsymbol{w}|\boldsymbol{x};\Lambda) = \frac{1}{Z(\Lambda)} P(\boldsymbol{x}|\boldsymbol{w})P(\boldsymbol{w}) \exp \sum_j \lambda_j f_j(\boldsymbol{w}), \quad (2)$$

where f_j denotes a feature function derived from a word or phoneme n-gram context, which returns the number of

sequence occurring in $\boldsymbol{w}$, and $\lambda_j \in \Lambda$ is a weighting factor. $Z(\Lambda)$ denotes a normalization factor. Discriminative language modeling is equivalent to estimating the weighting factors, Λ, from labeled and unlabeled training lattices when Eq. (2) is viewed as a discriminative model.

2.2. Semi-Supervised Language Modeling

Semi-supervised language modeling is aimed at increasing LM robustness by incorporating information from a large amount of unlabeled lattices. One of the key issues concerning semi-supervised training is how the contribution of unlabeled lattices is reflected in the LM estimated from the labeled ones. Although the semi-supervised training is typically formulated as an optimization problem consisting of two independent objectives for different training lattices, it is difficult to find the optimum that minimizes both objectives simultaneously. To address this issue, the LM is estimated by introducing an approach called "multi-objective optimization programming" (MOP) [9], which was successfully applied for automatic language identification by Yaman *et al*. [10]. In this approach, a set of compromise solutions, *i.e.* a Pareto set, is obtained by accepting trade-offs between the objectives. The most preferred solution can be selected among the set. It indicates that this approach makes it possible to select an LM appropriate to a target broadcast program from possible LMs.

The ε-constraint method (**EPS**) is employed to solve a MOP problem [9]. By this method, one objective function is converted to an inequality constraint, and the other is minimized under the constraint as follows:

$$\Lambda' = \arg\min_{\Lambda} \mathrm{L}(\Lambda) \ \text{ s.t. } \ \mathrm{U}(\Lambda) \leq \bar{\mathrm{U}}, \tag{3}$$

where $\bar{\mathrm{U}}$ is a precomputed upper-bound value that is 5 to 20% lower than the objective at $\Lambda = \mathbf{0}$. This optimization is typically solved by using an augmented Lagrangian with a quadratic penalty [11], which is given by

$$F(\Lambda, \kappa; \rho) = L(\Lambda) + \rho \left\langle \frac{\kappa}{2\rho} + \bar{U} - U(\Lambda) \right\rangle^2, \tag{4}$$

where κ is a Lagrange multiplier, and ρ is a penalty parameter. $\langle x \rangle$ denotes an operator, $\max\{x, 0\}$. Since a set of solutions depending on the configuration of the inequality constraint is obtained, the solution that minimizes WER for a development set can be thus selected.

The weighted-sum method (**WS**) is commonly used to solve the MOP problem because of its simpleness. The optimal solution, Λ', is given by

$$\Lambda' = \arg\min_{\Lambda} \left\{ \mu_{\mathrm{L}} L(\Lambda) + \mu_{\mathrm{U}} U(\Lambda) \right\}, \tag{5}$$

where μ_{L} and μ_{U} are weighting factors for individual objective functions. The drawback of this method is that it may fail to obtain an optimum when a set of Pareto solutions forms a non-convex set [12]. The ε-constraint method and the weighted-sum method are compared in the following section.

3. OBJECTIVE FUNCTIONS

3.1. Risk-Based Objectives

The remaining issue concerning semi-supervised language modeling is how to design objective functions. With the proposed method, the objective functions for labeled or unlabeled training lattices are derived from the Bayes risk (**Risk**) [13]. As the objective for the labeled lattices is represented as a special case of the objective for the unlabeled lattices, the objectives are naturally integrated in terms of word error minimization.

Given a training audio input, $\boldsymbol{x}_m^{(\ell)}$ $(m = 1, \ldots, M)$, an objective function based on the Bayes risk [5] is defined as

$$\mathrm{L}_1(\Lambda) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\boldsymbol{w} \in \mathcal{L}_m} R(\boldsymbol{w}_m^r, \boldsymbol{w}) P(\boldsymbol{w} | \boldsymbol{x}_m^{(\ell)}; \Lambda), \tag{6}$$

where $R(\boldsymbol{w}_m^r, \boldsymbol{w})$ is a cost (*e.g.* Levenshtein distance) defined between the reference, $\boldsymbol{w}_m^r$, and the hypothesis, $\boldsymbol{w}$, in the m-th training lattice, $\mathcal{L}_m$.

The above function is easily extended to deal with unlabeled lattices [8]. Given an input, $\boldsymbol{x}_n^{(u)}$ $(n = 1, \ldots N)$, the unsupervised version of the objective is given by

$$\begin{aligned} \mathrm{U}_1(\Lambda) \ = \ & \frac{1}{N} \sum_{n=1}^{N} \sum_{\boldsymbol{w} \in \mathcal{L}_n} P(\boldsymbol{w} | \boldsymbol{x}_n^{(u)}; \Lambda) \times \\ & \sum_{\boldsymbol{w}' \in \mathcal{L}_n} R(\boldsymbol{w}, \boldsymbol{w}') P(\boldsymbol{w}' | \boldsymbol{x}_n^{(u)}; \Lambda) \end{aligned} \tag{7}$$

In practice, the expected risk for a lattice is efficiently approximated by using edge-wise risks. At edge e, the edge-wise risk, $\zeta(e)$, is given by

$$\zeta(e) \equiv \sum_{e' \in \mathrm{overlap}(e)} \ell_{0\text{-}1}(e, e') p(e'), \tag{8}$$

where $p(e)$ is an edge posterior, and $\ell_{0\text{-}1}(e, e')$ is a cost function defined between overlapping edges. A simple binary function given by

$$\ell_{0\text{-}1}(e, e') \equiv \begin{cases} 0 & \text{if } \mathrm{label}(e) = \mathrm{label}(e') \\ 1 & \text{otherwise,} \end{cases} \tag{9}$$

is used. The approximate risk of the lattice is computed by the forward-backward algorithm by using the edge-wise risks. A detailed description of the approximation can be found in [8].

3.2. Conditional Log-likelihood Based Objectives

One of the conventional discriminative objective functions used for labeled lattices is based on the negative conditional log-likelihood (**CLL**) [4]. The objective is defined as

$$\mathrm{L}_2(\Lambda) = -\frac{1}{M} \sum_{m=1}^{M} \log P(\boldsymbol{w}_m^r | \boldsymbol{x}_m^{(\ell)}; \Lambda). \tag{10}$$

Table 1. *Feature Functions*

		#features
phoneme	bigrams	1.3k
	trigrams	12.9k
word	bigrams	731.9k
	trigrams	1859.6k

Table 2. *Evaluation Data*

	#utts	#words	PP	OOV(%)	WER(%)
Dev.	245	3.5k	125.7	1.5	23.0
Test	551	7.0k	139.4	1.3	22.3

This objective function is analogous that used in maximum mutual information (MMI) acoustic modeling.

In order to incorporate information from the unlabeled lattices into the CLL-based objective function, we use minimum entropy regularization introduced in [14]. The regularizer is defined as conditional entropy given $\boldsymbol{x}_n^{(u)}$,

$$\mathrm{U}_2(\Lambda) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{\boldsymbol{w}\in\mathcal{L}_n} P(\boldsymbol{w}|\boldsymbol{x}_n^{(u)};\Lambda)\log P(\boldsymbol{w}|\boldsymbol{x}_n^{(u)};\Lambda). \tag{11}$$

Note that the combination of Eqs. (10) and (11) is similar to semi-supervised acoustic modeling described in [2].

4. EXPERIMENTS

4.1. Setup

NHK's speech decoder transcribes audio streams of broadcast programs in real time, while detecting start and end points of speech segments [15]. 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 employs a two-pass strategy that obtains 200-best sentence hypotheses by using gender-dependent HMMs and a bigram LM in the first pass and rescores them using a trigram LM with scores derived from the weighting factors, Λ.

The acoustic models were obtained from a total of 650 hours of speech in broadcast news programs using minimum phone error (MPE) training [16]. The baseline trigram LM denoted as $P(\boldsymbol{w})$ in Eq. (2) was trained on Japanese broadcast news manuscripts and transcriptions (239M words), and the vocabulary size was set to 100k. The linguistic feature functions in Eq. (2) were defined by word or phoneme bigrams and trigrams observed more than five times in the labeled and unlabeled training lattices (Table 1). The phoneme-based features were extracted along with gender information from the phoneme sequences embedded in the training word lattices.

Table 2 lists the evaluation data, taken from three episodes of an NHK news program, including conversational speech and voice-overs. Two episodes were used as test data, and the remaining episode was used as development data.

Table 3. *Training Data*

	#hours	#utts	#words	WER(%)
Labeled	58.6	26k	697.5k	22.3
Unlabeled	344.1	218.6k	2.84M	n/a

Table 4. *Experimental Results*

			Dev	Test
Baseline			23.0	22.3
Mixture LM			22.8	22.3
Labeled (supervised)	CLL		22.9	22.1
	Risk		22.8	22.3
Unlabeled (unsupervised)	Entropy		22.7	22.2
	Risk		22.3	21.5
Labeled+Unlabeled (semi-supervised)	WS	CLL+Ent.	22.7	22.0
		Risk	22.0	21.2
	EPS	CLL+Ent.	22.5	22.0
		Risk	**21.9**	**20.9**

Table 3 shows the labeled or unlabeled training data. The labeled data (including conversational speeches on news topics) were taken from the same news program. The unlabeled data were taken from similar broadcast programs including conversational speeches, *e.g.* discussions and debates about current news topics. Semi-supervised language modeling was performed on the decoded lattices of these training data. Since there were no reference transcriptions available for the unlabeled data, the number of utterances and the number of words were quantified by the decoder in the table.

4.2. Results

Table 4 lists the WER results for the evaluation data. In the table, Baseline denotes the results from the baseline trigram LM without discriminative training. Two types of MOP methods for semi-supervised training (Labeled+Unlabeled) were used, and EPS represents the results obtained by the ε-constraint method (*cf.* Eq. (3)), while WS denotes results obtained by the weighted-sum method (*cf.* Eq. (5)). For comparison, the results obtained from linear-interpolated LMs (Mixture LM) are also shown. The LMs were interpolated between the baseline LM and the LM estimated from the ASR transcriptions of unlabeled data. The interpolation weights were estimated from the development data. The results obtained from the models trained by the labeled lattices in the supervised manner (Labeled) and those obtained from the unlabeled lattices in the unsupervised manner (Unlabeled) are also shown.

The results obtained from Mixture LM achieved small WER reductions for Baseline. The interpolated LMs would

limit the efficiency of WER reductions because the topics in the evaluation data are not covered by the unlabeled data. In supervised language modeling, CLL and Risk also achieved small WER reductions for Baseline. It is probable that there are too few training lattices for estimating the weighting factors with statistical reliability. In contrast to Labeled, Risk (Unlabeled), which was estimated by using the expected risks (*cf.* Eq. (7)), outperformed the model obtained from labeled training lattices. For the test data, it achieved a WER of 22.3% and produced relative reduction of 3.6% compared with Baseline. Since language modeling was carried out using unlabeled lattices that were over six times larger than the labeled lattices, a more robust model was obtained.

In the case of semi-supervised language modeling, all the results for the test data showed further reductions in WER compared with the results from Labeled and Unlabeled. Especially, Risk (EPS), which was trained by the ε-constraint method using risk-based objective functions, achieved WER of 20.9% for the test data and provided a relative reduction of 6.3% for Baseline and 2.8% for Risk (Unlabeled), respectively. According to a matched-pair test, WER was decreased at a significance level of 0.05. Although these results reveal that the MOP approaches are effective for reducing WER, no significant differences between the results obtained by the weighted-sum method (WS) and those obtained by the ε-constraint method were found. In contrast, the risk-based objective functions provided significant reductions in WER compared with the CLL-based objectives regardless of the MOP approach taken. This is because the risk-based objectives are closely associated with WERs, as is clear from their definitions denoted by Eqs. (6) and (7). The performance of semi-supervised language modeling depends on a variety of factors (such as formulations of objectives and sizes of the training lattices) rather than on the MOP approach taken. To design the efficient semi-supervised language modeling approach, a further experimental study is therefore required. From a qualitative point of view, the proposed MOP approach reduced more deletion and insertion errors by short words such as Japanese particles than Baseline and the CLL-based approaches.

5. CONCLUSION

A semi-supervised language modeling method, designed to improve robustness of a language model (LM) obtained from labeled data, is proposed. The LM is defined as a log-linear model, which employs a set of linguistic feature functions. Risk-based objective functions are derived from labeled or unlabeled training lattices, and the LM is given by a solution to the problem of multi-objective optimization programming. Experimental results showed that the MOP-based approach successfully integrates the objectives and significantly reduces word error rate in transcribing Japanese broadcast programs. In future work, the model will be modified to introduce regularizers into the risk-based objectives.

6. REFERENCES

[1] T. Imai, S. Homma, A. Kobayashi, T. Oku, and S. Sato, "Speech recognition with a seamlessly updated language model for real-time closed-captioning," in *Proc. Interspeech*, pp. 262–265, 2010.

[2] J.-T. Huang and M. Hasegawa-Johnson, "Semi-supervised training of gaussian mixture models by conditional entropy minimization," in *Proc. Interspeech*, pp. 1353–1356, 2010.

[3] X. Liu, J. Huang, and J.-T. Chien, "Multi-view and multi objective semi-supervised learning for large vocabulary continuous speech recognition," in *Proc. ICASSP*, pp. 4668–4671, 2011.

[4] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," *Computer Speech and Language*, vol. 21, pp. 373–392, 2007.

[5] A. Kobayashi, T. Oku, S. Homma, S. Sato, T. Imai, and T. Takagi, "Discriminative rescoring based on minimization of word errors for transcribing broadcast news," in *Proc. Interspeech*, pp. 1574–1577, 2008.

[6] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proc. ICASSP*, pp. 224–227, 2003.

[7] D. Mrva and P.C. Woodland, "Unsupervised language model adaptation for mandarin broadcast conversation transcription," in *Proc. Interspeech*, pp. 2210–2213, 2006.

[8] A. Kobayashi, T. Oku, S. Homma, T. Imai, and S. Nakagawa, "Lattice-based risk minimization training for unsupervised language model adaptation," in *Proc. Interspeech*, pp. 1453–1456, 2011.

[9] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, pp. 369–395, 2004.

[10] S. Yaman and C.-H. Lee, "A flexible classifier design framework based on multi-objective programming," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 779–789, 2008.

[11] J.A. Snyman, *Practical mathematical optimization*, Springer, 2005.

[12] K. Miettinen, "*Nonlinear multiobjective optimization*," Springer, 1999.

[13] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," *Computer Speech and Language*, vol. 14, pp. 115–135, 2000.

[14] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, pp. 529–536, 2005.

[15] T. Imai, S. Sato, A. Kobayashi, K. Onoe, and S. Homma, "Online speech detection and dual-gender speech recognition for captioning broadcast news," in *Proc. Interspeech*, pp. 1602–1605, 2006.

[16] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, pp. I–105–108, 2002.