

Lattice-Based Risk Minimization Training for Unsupervised Language Model Adaptation

Akio Kobayashi¹, Takahiro Oku¹, Shinichi Homma¹, Toru Imai¹ and Seiichi Nakagawa²

¹NHK Science and Technology Research Laboratories, Tokyo, Japan

²Toyohashi University of Technology, Toyohashi, Japan

{kobayashi.a-fs, oku.t-le, homma.s-fc, imai.t-mq}@nhk.or.jp, nakagawa@slp.cs.tut.ac.jp

Abstract

This paper describes a lattice-based risk minimization training method for unsupervised language model (LM) adaptation. In a broadcast archiving system, unsupervised LM adaptation using transcriptions generated by speech recognition is considered to be useful for improving the performance. However, conventional linear interpolation methods occasionally degrade the performance because of incorrect words in the training transcriptions. Accordingly, we propose a new adaptation method aiming to reflect error information among training lattices. The method minimizes the whole risk of training lattices to yield a log-linear model, which consists of a set of linguistic features. The advantage of the method is that the model parameters can be obtained efficiently in an unsupervised manner. Experimental results obtained in transcribing Japanese broadcast news showed significant word error rate reduction for those of conventional mixture LMs.

Index Terms: unsupervised learning, Bayes risk minimization, language model adaptation, speech recognition

1. Introduction

The recent progress being made in the field of corpus-based spoken-language processing has led to significantly successful applications in the real world. NHK (Japan Broadcasting Corp.) has developed a real-time automatic speech recognition (ASR) system for closed-captioning broadcast news [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 involving spoken term detection, document retrieval, etc. In the last decade, many ASR-driven broadcast archiving systems have been proposed [2, 3]. Recently, similar archiving systems have been utilized not merely for broadcast news programs but for meetings where a variety of spontaneous speech phenomena are observed [4].

In the archiving systems, unsupervised training or adaptation could be effective to make upcoming transcriptions more accurate, because an increasing amount of contents such as ASR transcriptions becomes available each and every second. In the perspective of language model (LM) adaptation, unsupervised adaptation is typically conducted by n-gram count merge or model-based linear interpolation methods [5, 6, 7]. However, these methods do not always perform the best in terms of word error rates (WERS), since the LMs are estimated by using ASR transcriptions containing misrecognized incorrect words as training data. Obviously, adjustments to the LMs are needed to reduce the influence of these erroneous words.

To this end, we introduce a lattice-based risk minimization training (RMT) method for unsupervised LM adaptation, which

aims to reflect error information in training data. The most distinctive feature of the proposed training method is that it enables efficient computation of LM adaptation by using lattices as training data in an unsupervised manner. In the method, a log-linear model is estimated to minimize the whole risk, which is similar to the risk defined in the conventional Bayes risk minimization approach [8, 9]. However, in contrast to the conventional approach, hypothesis selection is made by use of a statistical model, whose parameters are estimated from lattices.

This paper describes a computational method for deriving the risk from lattices and its derivatives to solve the problem using gradient-based algorithms in detail.

2. Lattice-Based Risk Minimization for Unsupervised LM Adaptation

2.1. Bayes Risk Minimization

The Bayes risk minimization approach is conventionally performed on n-best lists to obtain hypotheses with minimum error probabilities [8, 9].

$$\hat{w} = \arg \min_w \sum_{w'} \ell(w, w') P(w'|x), \quad (1)$$

where $P(w'|x)$ is a posterior probability for a sentence hypothesis w' derived from an n-best list for an audio input x and $\ell(w, w')$ is a cost function defined as a Levenshtein distance between two hypotheses. Although this approach is applied to individual n-best lists in an unsupervised manner, it is computationally expensive due to the need to evaluate every pair of hypotheses in Eq. (1).

2.2. Lattice-Based Risk Minimization Training for Unsupervised LM Adaptation

The risk minimization problem in Eq. (1) can be solved by the inductive learning strategy, and a variety of studies performed on the basis of the Bayes decision rule have appeared in the literature [10, 11, 12]. In [10], a discriminative approach that minimizes the whole risk of training lattices in a supervised manner was proposed to aim at reflecting error information in training data. On the basis of a similar motivation, we explored a discriminative approach to language modeling on lattices in a supervised manner [12]. One of the advantages of using lattices is that the training method can efficiently utilize all the information about hypotheses because of their compact graph representations. In addition, extra parameters such as n-best list sizes are no longer needed. However, lattice-based RMT in an unsupervised manner has not been studied well so far. Therefore, we propose a new scheme that minimizes the risk of training

lattices for unsupervised LM adaptation by taking the approach in [12] one step further.

Initially, we define an unsupervised version of the whole risk of training data so that it reflects relevant information about word errors. Given an utterance, \mathbf{x}_m ($m = 1, \dots, M$), and the n -th corresponding sentence hypothesis, $\mathbf{w}_{m,n}$, the whole risk is given by

$$\mathcal{R}(\Lambda) = \frac{1}{M} \sum_m \sum_n P(\mathbf{w}_{m,n}|\mathbf{x}_m; \Lambda) \gamma(\mathbf{w}_{m,n}), \quad (2)$$

where $P(\mathbf{w}_{m,n}|\mathbf{x}_m; \Lambda)$ is a posterior probability modeled by a log-linear form,

$$P(\mathbf{w}_{m,n}|\mathbf{x}_m; \Lambda) = \frac{1}{Z(\Lambda)} P(\mathbf{w}_{m,n}|\mathbf{x}_m) \exp \sum_j \lambda_j f_j(\mathbf{w}_{m,n}). \quad (3)$$

In this paper, f_j denotes a word n-gram feature function that returns the number of n-grams occurred in $\mathbf{w}_{m,n}$ and $\lambda_j \in \Lambda$ is a corresponding weight. Minimizing the risk leads to a log-linear model that reflects tendencies of misrecognized incorrect words in the training data. $\gamma(\mathbf{w}_{m,n})$ in Eq. (2) denotes a loss function, which is given by

$$\gamma(\mathbf{w}_{m,n}) = \sum_{n'} \ell(\mathbf{w}_{m,n}, \mathbf{w}_{m,n'}) P(\mathbf{w}_{m,n'}|\mathbf{x}_m; \Lambda). \quad (4)$$

Since Eq. (4) is defined among sentence hypotheses, we need to re-define the risk computable on the training lattices. Thus, accumulating local edge-wise risks from the bottom up, we obtain the whole risk of the lattice in a similar manner to minimum phone error (MPE) training [13]. Then, we introduce an edge-wise risk function, which is substituted for the Levenshtein distance originally used in the Bayes risk minimization. Given a lattice, \mathcal{L}_m , derived from the m -th utterance, and overlapping edges, e and e' , we define an edge-wise cost function, $\ell_{0-1}(e, e')$, as follows:

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

An edge-wise risk, which roughly represents the local amount of information about word errors, is defined as

$$\zeta(e) \equiv \sum_{e' \in \text{overlap}(e)} \ell_{0-1}(e, e') p(e'), \quad (5)$$

where $p(e')$ is an edge posterior probability given by

$$p(e) = \frac{1}{\alpha} \{\alpha(\sigma(e)) \cdot g(e) \cdot \beta(\tau(e))\}, \quad (6)$$

where $\sigma(e)$ denotes a start node for e , while $\tau(e)$ represents an end node. $\alpha(\sigma(e))$ is a forward probability at $\sigma(e)$, and $\bar{\alpha}$ denotes that of the final node. $\beta(\tau(e))$ is a backward probability at $\tau(e)$. $g(e)$ is a transition score computed from an acoustic model score, $\varphi_{am}(e)$, a language model score, $\varphi_{lm}(e)$, and scores given by weighting factors in Eq. (3) as follows:

$$g(e) = e^{\lambda_{am} \varphi_{am}(e) + \lambda_{lm} \varphi_{lm}(e) + \sum_j \lambda_j \varphi_j(e)}, \quad (7)$$

where constants, λ_{am} and λ_{lm} , are scaling factors, respectively. A binary function, $\varphi_j(e)$, returns 1 if f_j is activated on e .

According to the forward algorithm, the cumulative risk, $q(t)$, at the node t is computed by

$$q(t) = \frac{\sum_{e: \tau(e)=t} \{g(\sigma(e)) + \zeta(e)\} \{\alpha(\sigma(e)) \times g(e)\}}{\alpha(t)}. \quad (8)$$

Applying Eq. (8) topologically from the initial node to the final node of the lattice, we obtain the forward risk, Υ_m . Consequently, the whole risk of the training data, \mathcal{R} , is given by $1/M \sum_m \Upsilon_m$.

The weighting factors, Λ , are derived by solving the minimization problem for the risk of the training lattices. Here, we solve this problem with quasi-Newton or gradient-based methods and develop approximations of gradients of \mathcal{R} with regard to Λ .

In the beginning, the backward risk, $r(\tau(e))$, is computed according to the backward algorithm, and the expected risk of all the paths passing through e is calculated by

$$v(e) = q(\sigma(e)) + \zeta(e) + r(\tau(e)). \quad (9)$$

On the other hand, Υ_m can be decomposed as

$$\Upsilon_m = p(e)v(e) + (1 - p(e))v'(e), \quad (10)$$

where $v'(e)$ denotes the risk of all the paths not passing through e . Assuming that the weighting factors, Λ , depend only on the edge posteriors, $p(e)$, and $v(e)$ and $v'(e)$ are regarded as constants, the following approximate derivatives can be obtained.

$$\frac{\partial \Upsilon_m}{\partial p(e)} \approx v(e) - v'(e), \quad (11)$$

$$\frac{\partial p(e)}{\partial \lambda_j} \approx p(e)(1 - p(e)) \varphi_j(e). \quad (12)$$

The gradient, $\delta_{j,e}^m$, at e is approximately computed by

$$\delta_{j,e}^m = -p(e)(v(e) - \Upsilon_m) \varphi_j(e). \quad (13)$$

The gradient of the m -th lattice for λ_j is given by $\sum_{e \in \mathcal{L}_m} \delta_{j,e}^m$, and consequently the final gradient, Δ_j , is calculated by

$$\Delta_j = \frac{1}{M} \sum_m \sum_{e \in \mathcal{L}_m} \delta_{j,e}^m. \quad (14)$$

Note that, in decoding, the score of a sentence hypothesis \mathbf{w} is given by the logarithmic acoustic and language model scores plus $\sum_j \lambda_j f_j(\mathbf{w})$.

2.3. Risk Minimization Training Using N-Best Lists

The objective function can be also minimized using n-best lists instead of lattices. In this case, the gradients are straightforwardly calculated from Eq. (2) and Eq. (3) as follows:

$$\Delta_j = -\frac{1}{M} \sum_m \sum_n P(\mathbf{w}_{m,n}|\mathbf{x}_m; \Lambda) \rho_j(\mathbf{w}_{m,n}), \quad (15)$$

where $\rho_j(\mathbf{w}_{m,n})$ is given by

$$\rho_j(\mathbf{w}_{m,n}) = \sum_{n'} \ell(\mathbf{w}_{m,n}, \mathbf{w}_{m,n'}) P(\mathbf{w}_{m,n'}|\mathbf{x}_m; \Lambda) \cdot \left[f_j(\mathbf{w}_{m,n}) - \sum_t f_j(\mathbf{w}_{m,t}) P(\mathbf{w}_{m,t}|\mathbf{x}_m; \Lambda) + f_j(\mathbf{w}_{m,n'}) - \sum_t f_j(\mathbf{w}_{m,t}) P(\mathbf{w}_{m,t}|\mathbf{x}_m; \Lambda) \right]. \quad (16)$$

Note that we use the partial derivative of Eq. (3) with regard to λ_j , $\partial P(\mathbf{w}_{m,n}|\mathbf{x}_m; \Lambda) / \partial \lambda_j = P(\mathbf{w}_{m,n}|\mathbf{x}_m; \Lambda) \{f_j(\mathbf{w}_{m,n}) - \sum_t f_j(\mathbf{w}_{m,t}) P(\mathbf{w}_{m,t}|\mathbf{x}_m; \Lambda)\}$. Since Eq. (15) is completely different from Eq. (14), the contrasting descent directions would lead to different sets of weighting factors. Therefore, these two types of approaches are compared in the following section.

Table 1: *Development and Evaluation Data (Sports News)*

	#utts	#words	PP	OOV(%)	WER(%)
Dev.	441	3.7k	210.7	1.3	40.3
Eval.	1387	11.1k	210.3	2.0	45.0

Table 2: *Development and Evaluation Data (Economic News)*

	#utts	#words	PP	OOV(%)	WER(%)
Dev.	196	3.0k	152.0	1.2	25.3
Eval.	783	8.9k	151.6	1.3	26.4

3. Experiments

3.1. System

We briefly describe the speech recognizer with which the NHK’s broadcast archiving system is equipped. The system collects transcriptions of NHK’s broadcast news programs with corresponding video/audio streams, and the recognizer decodes the captured audio streams in real-time, while detecting start and end points of speech segments [14]. 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 recognizer employs a 2-pass strategy that obtains 200-best sentence hypotheses using gender dependent tree lexica and a bigram LM in the first pass and rescores them using a trigram LM with $\sum_j \lambda_j f_j(\mathbf{w})$ obtained by the RMT.

3.2. Experimental Setup

As listed in Tables 1 and 2, we prepared two different categories of experimental data, and each category has four individual news shows: one for development and three for evaluation. The perplexities (PP), out of vocabulary (OOV) rates, and word error rates (WER) were measured by a baseline trigram LM. The data in Table 1 contain sports news shows and are characterized by the topics including sumo tournaments, professional baseball games, and domestic league soccer games. The existence of background music also distinguishes the shows from others. The data in Table 2 consist of economic news shows, and each show has expert interviews and talks by financial analysts and economists in a conversational manner. The acoustic models were obtained from a total of 650 hours of speech in broadcast news shows using MPE training [13]. The baseline trigram LM was trained on Japanese broadcast news manuscripts and transcriptions (239M words), and the vocabulary size was set to 100k. The mixture LMs used in the experiments were obtained from model-based linear interpolation between the baseline LM and additional LMs from the portions of in-domain 1-best ASR transcriptions. Table 3 shows the training data for construction of the additional LMs. For each evaluation set, we prepared a total of 40 or 70 hours of ASR transcriptions as program-dependent or in-domain training data, respectively. Note that the number of utterances and that of words were quantified by ASR, since there were no reference transcriptions available on the training data. The word hypotheses in the transcriptions were filtered by threshold values determined by the word posterior probabilities. The values were obtained by evaluating the development sets. The mixture weights were also estimated by using the development sets.

The data in Table 3 were also used in the RMT. Trigram lattices generated by ASR were employed in the lattice-based

Table 3: *Training Data*

	#programs	hours	#segments	#words
Sports	143	70.0	72.5k	61.9k
Economic	138	40.0	26.4k	43.8k

Table 4: *WERs for RMT with Baseline LM*

	Training (hours)	Dev (%)		Eval (%)	
		n-best	lattice	n-best	lattice
Sports	0.0 (baseline)	40.3		45.0	
	5.0	39.1	39.1	44.2	44.0
	70.0(all)	39.2	39.1	43.6	43.5
Economic	0.0 (baseline)	25.3		26.4	
	5.0	23.6	23.5	25.4	25.4
	40.0(all)	23.9	23.4	25.6	25.4

RMT, whereas 200-best lists taken from the lattices were used in the RMT with n-best lists. The feature functions were defined by word bigrams and trigrams observed more than five times in the training data. The weighting factors were obtained by the L-BFGS algorithm [15]. The iteration counts of the RMT were determined by evaluating the development sets.

3.3. Experimental Results

3.3.1. Performance of Lattice-Based RMT

Initially, we examined the performance of lattice-based RMT compared to that of RMT using n-best lists as training data. In this experiment, however, we used the baseline LM instead of mixture LMs, aiming for evaluation of the pure RMT performance without any biases to LMs. As shown in Table 4, the lattice-based RMT achieved equal or superior results compared to those from the RMT with n-best lists. Since the latter can be carried out without any approximations, these results reveal that the approximate computation of the lattice-based RMT basically performed well, namely, the risk minimization could be conducted on the training lattices in an unsupervised manner. However, further detailed analyses are required from the theoretical and empirical viewpoint of the designs of edge-wise risks and the approximate gradients.

Table 4 also shows the WERs with different amounts of training transcriptions. In the sports news evaluation set, the WERs were decreased from 44.0 % to 43.5 % by the lattice-based RMT as well as from 44.2 % to 43.6 % by the RMT with n-best lists when the amount of transcriptions was increased. In contrast, the WERs were equal or inferior in the economic news evaluation set, since the total amount of training data was smaller than that of the sports news. This indicates that a substantial amount of training data would be needed to achieve significant WER improvement. Hence, we will examine the impact of training data by using even larger amounts of data in future work.

Compared with the baseline LM, the results from lattice-based RMT achieved a relative reduction of 3.3 % in the sports evaluation set (from 45.0 % to 43.5 %), and 3.8 % in the economic news evaluation set (from 26.4 % to 25.4 %).

3.3.2. Lattice-Based RMT with Mixture LM

We then investigated the performance of lattice-based RMT with mixture LMs. Table 5 shows the results evaluated for dif-

Table 5: WERs for Lattice-Based RMT with Mixture LMs

	Training (hours)	Dev (%)		Eval (%)	
		Mix	+RMT	Mix	+RMT
Sports	5.0	37.5	36.9	41.6	40.8
	70.0(all)	33.3	32.8	38.5	38.1
Economic	5.0	24.1	23.6	24.8	24.8
	40.0(all)	22.6	22.1	24.0	23.7

ferent amounts of training data. Note that the threshold value for filtering the word hypotheses was set to 0.5 in the table. The results from mixture LMs listed in the columns labeled "Mix" show larger WER reduction than those from the RMT which was applied exclusively (*cf.* Table 4) against the baseline results. As shown in the columns labeled "+RMT", the lattice-based RMT significantly reduced the WERs compared with the results from mixture LMs when the RMT was conducted on all the training lattices. The result for the sports news evaluation set achieved a WER of 38.1 % and produced a relative reduction of 1.0 %. Similarly, the result for the economic news was 23.7 % and a relative reduction of 1.3 %. According to a matched pair test [16], both WERs were decreased at a significance level of 0.05.

Finally, in order to investigate the efficiency of the amounts of data, we examined the lattice-based RMT for the development set of the sports news shows with various amounts of training data. Figure 1 plots the WERs obtained by the mixture LMs and those obtained by the lattice-based RMT along with the mixture LMs. The thresholds for filtering word hypotheses were varied among the values 0.0, 0.5, and 0.8 as well. At all the plot points (with one exception), the lattice-based RMT achieved further reductions over the results when only the mixture LMs were used. The reductions were small compared with those obtained by the use of lattice-based RMT without mixture LMs, because the mixture LMs fixed a lot of correctable errors instead of the lattice-based RMT. Therefore, in addition to the linguistic features such as word n-grams, many different types of features derived from acoustic feature vectors and phonetic sequences would be required to reduce WERs much further.

4. Conclusions

We proposed a lattice-based risk minimization training (RMT) method for unsupervised language model adaptation. We

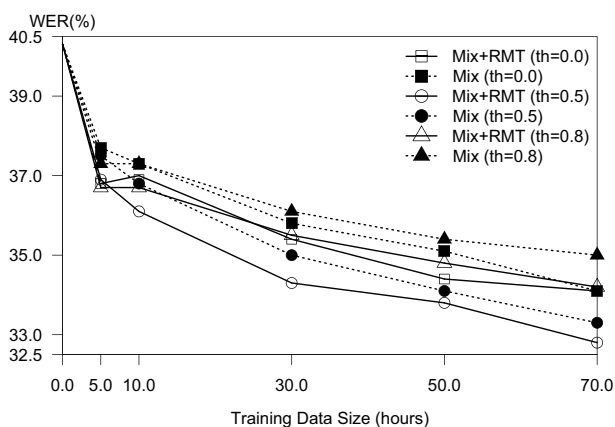


Figure 1: Lattice-Based RMT with Various Amounts of Training Data (Sports News, Development Set.)

demonstrated the derivation of the whole risk for training lattices and its derivatives using several approximations. The advantage of lattice-based RMT is that it makes more efficient use of information with respect to hypotheses on the lattices than similar training methods on n-best lists. Experimental results showed that the lattice-based RMT achieved significant WER reductions for the evaluation of news shows as compared to the results obtained with conventional linear-interpolated mixture LMs. The lattice-based RMT proved to be effective with statistical significance when large amounts of training transcriptions were available. In future work, we intend to take acoustic and phonetic features into consideration for further improvement in unsupervised adaptation.

5. 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] A. Hauptmann and M. Smith, "Text, speech, and vision for video segmentation: The Informedia project," in *Proc. AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*, pp. 90–95, 1995.
- [3] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," in *Proceedings of the IEEE*, vol. 88, pp. 1338–1353, 2000.
- [4] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*, pp. IV–357–360, 2007.
- [5] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Proc. ICASSP*, pp. 4297–4300, 2009.
- [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] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," *Computer Speech and Language*, vol. 14, pp. 115–135, 2000.
- [9] R. Schlüter, M. Nussbaum-Thom, and H. Ney, "On the relation of Bayes risk, word error, and word posteriors in ASR," in *Proc. Interspeech*, pp. 230–233, 2010.
- [10] V. Doumliotis, S. Tsakalidis, and W. Byrne, "Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition," *Speech Communication*, vol. 48, no. 2, pp. 142–160, 2006.
- [11] S.-H. Lin and B. Chen, "A risk minimization framework for extractive speech summarization," in *Proc. the 48th Annual Meeting of the ACL*, pp. 79–87, 2010.
- [12] 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.
- [13] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, pp. I–105–108, 2002.
- [14] 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.
- [15] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Programming*, vol. 45, no. 3, pp. 503–528, 1989.
- [16] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, pp. 532–535, 1989.