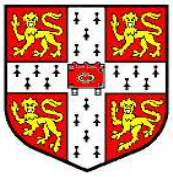

CAMBRIDGE UNIVERSITY

ENGINEERING DEPARTMENT

DISCRIMINATIVE LANGUAGE MODEL ADAPTATION FOR MANDARIN BROADCAST SPEECH TRANSCRIPTION AND TRANSLATION

X. A. Liu, W. J. Byrne, M. J. F. Gales & P. C. Woodland

CUED/F-INFENG/TR.586

Sept 2007

Cambridge University Engineering Department
Trumpington Street
Cambridge. CB2 1PZ
England

E-mail: xl207@eng.cam.ac.uk
http://mi.eng.cam.ac.uk/~xl207

Abstract

This paper investigates unsupervised test-time adaptation of language models (LM) using discriminative methods for a Mandarin broadcast speech transcription and translation task. A standard approach to adapt interpolated language models to is to optimize the component weights by minimizing the perplexity on supervision data. This is a widely made approximation for language modeling in automatic speech recognition (ASR) systems. For speech translation tasks, it is unclear whether a strong correlation still exists between perplexity and various forms of error cost functions in recognition and translation stages. The proposed minimum Bayes risk (MBR) based approach provides a flexible framework for unsupervised LM adaptation. It generalizes to a variety of forms of recognition and translation error metrics. LM adaptation is performed at the audio document level using either the character error rate (CER), or translation edit rate (TER) as the cost function. An efficient parameter estimation scheme using the extended Baum-Welch (EBW) algorithm is proposed. Experimental results on a state-of-the-art speech recognition and translation system are presented. The MBR adapted language models gave the best recognition and translation performance and reduced the TER score by up to 0.54% absolute.

1 Introduction

A crucial component in both an automatic speech recognition system and a statistical machine translation system is the language model. In order to more robustly handle different styles or tasks, LM adaptation schemes may be required. Due to data sparsity, directly adapting N-gram word probabilities is non-trivial. A standard approach is to re-adjust the interpolation weights of a mixture model by minimizing the perplexity on some supervision data. An assumption is made that there is a strong correlation between perplexity and error rate [1]. It is believed to be a good approximation to word error rate (WER) and widely used in current ASR systems [11].

However, for speech translation tasks such approximation can be poor. First, for logogram based languages such as Mandarin Chinese, there are no natural word boundaries in normal texts. Recognition performance is normally evaluated using character error rate. A widely adopted approach is to partition a string of characters into a sequence of "words". Language models are then trained on the resulting tokenized texts [12]. Due to the ambiguity in this character to word decomposition process, it may be argued that word level perplexity reduction may not necessarily lead to CER improvement. Secondly, performance of current SMT systems is typically measured in BLEU [2], or the translation edit rate (TER) metric [3]. It is also unclear whether a strong correlation exists between perplexity and translation error metrics.

One approach to address this issue is to use discriminative training techniques. These schemes do not make incorrect modeling assumption and explicitly aim at reducing the recognition, or translation, error rate. Along this line there has been research interest in discriminatively training parameters of N-gram language models for speech recognition [18, 21], and LM adaptation for SMT systems [8, 16]. Good performance improvements have been reported. Nonetheless, these current approaches are restricted to a certain form of cost function, and heavily rely on numerical methods during parametric optimization. Hence for complicated tasks like speech translation it would be interesting to employ a more flexible discriminative scheme that can generalize to various forms of error metrics at different stages of the system, which also has an efficient parametric optimization method. One such scheme is minimum Bayes risk (MBR) training [5, 6]. It has been successfully applied to speech recognition and can generalize to a variety forms of error cost functions.

This paper investigates using the MBR criterion for unsupervised discriminative language model adaptation in test-time for speech recognition and translation systems. LM adaptation is performed at the audio document level. Two forms of error metrics are used in MBR adaptation: the character error rate for speech recognition; the translation edit rate for later translation of the ASR output. The rest of the paper is organized as follows. Section 2 introduces linear and log-linear interpolations for mixture language models and reviews standard maximum likelihood based adaptation schemes. Section 3 introduces the MBR criterion and details the algorithms for discriminatively adapting LM interpolation weights in both linear and log-linear cases. An efficient re-estimation scheme based on the extended Baum-Welch (EBW) algorithm is presented. In section 4 a number of implementation issues are discussed. In section 5 experimental results on a stat-of-the-art Mandarin broadcast speech transcription and translation system are presented. Section 6 is the conclusion and discussion of future work.

2 Maximum Likelihood LM Adaptation

A common form of a mixture language model is to interpolate word probabilities using linear weights. For N-gram word based models considered in this paper, this is given by,

$$P(w_i|h_{i-N+1}^{i-1}) = \sum_m \lambda_m P_m(w_i|h_{i-N+1}^{i-1}) \quad (1)$$

where w_i denote the i word of a word sequence, $\mathcal{W}$, h_{i-N+1}^{i-1} its N-gram history, and λ_m, the interpolation weight for the mth component model, $P_m(\cdot)$.

Alternatively word probabilities may be linearly interpolated in the log space,

$$P(w_i|h_{i-N+1}^{i-1}) = \frac{1}{Z} \exp\left(\sum_m \lambda_m \log P_m(w_i|h_{i-N+1}^{i-1})\right) \quad (2)$$

where Z is a normalization term to ensure the interpolated probability to be a valid distribution. As the weights are applied directly to the log-likelihood scores of individual LM components, such a model may

provide more power to capture the curvature of the likelihood function. It my be related to a multiple stream HMM system using different front-end processing schemes, or the log-interpolation of "feature functions" in SMT systems [8].

One issue with a log-linear model is that the exact calculation of the normalization term is non-trivial. Hence it is difficult to give a probabilistic interpretation and derive the required likelihood based estimation scheme. For the same reason, when applying these models in a full search on ASR or SMT tasks, there is a lack of efficient back-off schemes which requires all interpolated N-gram probabilities are valid distributions. However, this may not be an issue for discriminative methods or posterior based techniques as the normalization term often may be canceled out [18]. This will be further discussed later for MBR adaptation. The rest of this section focuses on likelihood based adaptation for linear interpolated models.

PP based adaptation: The interpolation weights are re-estimated to minimize the perplexity on hypotheses generated from a previous pass of an ASR or SMT system. This is equivalent to maximizing the joint probability of the entire word sequence in the supervision hypothesis. Take a mixture LM used in an ASR system as an example. Let $\hat{\mathcal{W}}$ denote the 1-best recognition hypothesis for a sequence of speech observations, $\mathcal{O}$. The optimal linear interpolation weight, λ_m, for the mth component model, $P_m(\cdot)$, can be derived by [1],

$$\begin{aligned}\hat{\lambda}_m &= \arg\max_{\lambda_m}\{\mathcal{F}_{\mathsf{ML}}(\mathcal{O})\} \\ &= \arg\max_{\lambda_m}\left\{\log p(\mathcal{O}|\hat{\mathcal{W}})P(\hat{\mathcal{W}})\right\}\end{aligned} \tag{3}$$

The acoustic distribution, $p(\mathcal{O}|\hat{\mathcal{W}})$ is independent of the language model parameters and therefore can be ignored. Assuming that $0 < \lambda_m < 1$ and $\sum_m \lambda_m = 1$, the Baum-Welch (BW) algorithm may be used to iteratively re-estimate the weights,

$$\hat{\lambda}_m = \frac{\tilde{\lambda}_m \left.\frac{\partial \mathcal{F}_{\mathsf{ML}}(\mathcal{O})}{\partial \lambda_m}\right|_{\lambda_m=\tilde{\lambda}_m}}{\sum_m \tilde{\lambda}_m \left.\frac{\partial \mathcal{F}_{\mathsf{ML}}(\mathcal{O})}{\partial \lambda_m}\right|_{\lambda_m=\tilde{\lambda}_m}} \tag{4}$$

where $\tilde{\lambda}_m$ is the current estimate of λ_m, and

$$\frac{\partial \mathcal{F}_{\mathsf{ML}}(\mathcal{O})}{\partial \lambda_m} = \sum_i \frac{P_m(w_i|h_{i-N+1}^{i-1})}{\sum_m \lambda_m P_m(w_i|h_{i-N+1}^{i-1})} \tag{5}$$

If perplexity base adaptation is performed in supervised mode the correct transcription is required.

Lattice/N-best based adaptation: As the error rate of the initial hypothesis increases, it becomes more useful to extend the above single hypothesis based adaptation to a lattice or N-best based approach. Rather than maximizing the likelihood of one reference, the marginal probability over multiple hypotheses, $\{\mathcal{W}\}$, is optimized,

$$\begin{aligned}\hat{\lambda}_m &= \arg\max_{\lambda_m}\{\mathcal{F}_{\mathsf{LAT}}(\mathcal{O})\} \\ &= \arg\max_{\lambda_m}\left\{\log\sum_{\mathcal{W}} p(\mathcal{O}|\mathcal{W})P(\mathcal{W})\right\}\end{aligned} \tag{6}$$

This technique has been widely used in unsupervised adaptation for acoustic models in state-of-the-art ASR systems [11]. The BW algorithm may still be used for lattice adaptation of LM weights. The sufficient derivative statistics required in the BW algorithm of equation 4 will be summed over all hypothesis and weighted by their posterior probabilities, $P(\mathcal{W}|\mathcal{O})$,

$$\frac{\partial \mathcal{F}_{\mathsf{LAT}}(\mathcal{O})}{\partial \lambda_m} = \sum_{\mathcal{W},i} P(\mathcal{W}|\mathcal{O}) \frac{P_m(w_i|h_{i-N+1}^{i-1})}{\sum_m \lambda_m P_m(w_i|h_{i-N+1}^{i-1})} \tag{7}$$

Posterior adaptation: Insufficient supervision data may lead to unrobust model adaptation. One approach to address such parametric uncertainty is to use posterior adaptation. Rather than directly optimize the interpolation weights, their prior distribution and the associated hyper-parameters are optimized. $\{\psi_m\}$ are optimized,

$$\begin{aligned}\hat{\psi}_m &= \arg\max_{\psi_m}\{\mathcal{F}_{\mathsf{POST}}(\mathcal{O})\} \\ &= \arg\max_{\psi_m}\left\{\int p(\mathcal{O}|\mathcal{W})P(\mathcal{W})P(\lambda_m|\psi_m)d\lambda_m\right\}\end{aligned} \tag{8}$$

In this paper during LM adaptation the supervision data assumed to be sufficient. Hence posterior adaptation is not considered.

Now consider an analogy between ASR and SMT systems. An SMT system may also be partitioned into two distinctive components, the translation model, and the target language model. The translation model can be viewed as a generative distribution that produces the source language sentence from the target language translation. Under this analogy, the above likelihood based schemes may also be applied to LM adaptation for SMT. This simply requires replacing the recognition hypothesis by the translation hypothesis and the speech input by the ASR output. In the rest of this paper detailed derivations of discriminative LM adaptation will be presented in the context of ASR systems for brevity.

3 Minimum Bayes Risk LM Adaptation

The expected recognition error of an ASR system for a sequence of speech observations, $\mathcal{O}$, can be expressed as a sum over the performance contribution from all possible hypotheses $\{\mathcal{W}\}$, further weighted by their posterior probabilities, $P(\mathcal{W}|\mathcal{O})$. Hence the weight parameters are optimized by [5, 6],

$$\begin{aligned}\hat{\lambda}_m &= \arg\min_{\lambda_m}\{\mathcal{F}_{\mathsf{MBR}}(\mathcal{O})\} \\ &= \arg\min_{\lambda_m}\left\{\sum_{\mathcal{W}} P(\mathcal{W}|\mathcal{O})\mathcal{L}(\mathcal{W},\tilde{\mathcal{W}})\right\}\end{aligned} \tag{9}$$

where $\mathcal{L}(\mathcal{W},\tilde{\mathcal{W}})$ denotes the defined recognition error rate measure of hypothesis $\mathcal{W}$ against the reference hypothesis $\tilde{\mathcal{W}}$. Various forms of cost function, such as CER, may be used depending on the evaluation metric being considered. This provides more flexibility, compared with other discriminative criteria, such as maximum mutual information (MMI), as the cost function is not necessarily restricted to one particular form. By definition if $\tilde{\mathcal{W}}$ is the correct transcription MBR adaptation will be performed in supervised mode.

In this paper the cost function considered for SMT systems is the translation edit rate. The TER metric measures the ratio of the number of string edits between the target language hypothesis $\tilde{\mathbf{e}}$ and the reference translation $\mathbf{e}$ to the total number of words in the reference. The allowable edit types include substitutions, insertions, deletions and phrasal level shifts,

$$\mathcal{L}_{\mathsf{TER}}(\tilde{\mathbf{e}},\mathbf{e}) = \frac{\mathsf{Ins}+\mathsf{Del}+\mathsf{Sub}+\mathsf{Shft}}{L}\times 100\% \tag{10}$$

where L is the total number of words in the reference. The TER metric has been found a closer approximation to human evaluation of translation quality than purely precision based cost functions such as BLEU [3]. If phrasal shifts are not permitted, the TER metric simplifies to the well-known word error rate (WER) measure.

Numerical methods may be used to optimize the MBR criterion. However, these schemes can be slow and difficult to guarantee convergence. The Extended Baum-Welch (EBW) algorithm [9] provides an efficient iterative optimization scheme for a family of rational objective functions, including MBR, that can be expressed as the ratio of two rational polynomials with

- non-negative coefficients, and non-negative variables;
- all variables subject to a sum-to-one constraint.

For a set of free parameters of the non-negative and sum-to-one constraint, the re-estimation formulae is given by,

$$\hat{\lambda}_m = \frac{\tilde{\lambda}_m\left(\frac{\partial\mathcal{F}_{\mathsf{MBR}}(\mathcal{O})}{\partial\lambda_m}\Big|_{\lambda_m=\tilde{\lambda}_m}+D\right)}{\sum_m\tilde{\lambda}_m\left(\frac{\partial\mathcal{F}_{\mathsf{MBR}}(\mathcal{O})}{\partial\lambda_m}\Big|_{\lambda_m=\tilde{\lambda}_m}+D\right)} \tag{11}$$

where $\tilde{\lambda}_m$ is the current estimate of λ_m, and D is a tunable regularization constant controlling the convergence speed. This is exactly the case of training discrete parameters like language model interpolation weights.

In the rest of this section detailed weights updating schemes based on the EBW algorithm are presented for both linear and log-linear interpolated models. In both cases the weights are constrained to be positive and sum-to-one.

Linear Interpolation: As discussed, the EBW re-estimation formulae given in equation 11 can be used to estimate $\{\lambda_m\}$. This requires the computation of, $\partial \mathcal{F}_{\mathsf{MBR}}(\mathcal{O})/\partial\lambda_m$, the partial derivative of the expected recognition accuracy against the mth component model's weight, λ_m. Following the MBR criterion given in equation 9 and applying chains rule, this may be re-expressed as,

$$\frac{\partial \mathcal{F}_{\mathsf{MBR}}(\mathcal{O})}{\partial \lambda_m} = \sum_{\mathcal{W}} \frac{\partial P(\mathcal{W}|\mathcal{O})\mathcal{L}(\mathcal{W},\tilde{\mathcal{W}})}{\partial \log p(\mathcal{O},\mathcal{W})} \frac{\partial \log p(\mathcal{O},\mathcal{W})}{\partial \lambda_m} \tag{12}$$

where the first term can be derived as the following,

$$\frac{\partial P(\mathcal{W}|\mathcal{O})\mathcal{L}(\mathcal{W},\tilde{\mathcal{W}})}{\partial \log p(\mathcal{O},\mathcal{W})} = P(\mathcal{W}|\mathcal{O})\left[1 - P(\mathcal{W}|\mathcal{O})\right]\mathcal{L}(\mathcal{W},\tilde{\mathcal{W}}) \tag{13}$$

The second term is independent of the acoustic model distribution $p(\mathcal{O}|\mathcal{W})$, and effectively identical to the sufficient statistics required by the standard perplexity based weights optimization scheme given in equation 5.

Log-linear Interpolation: As discussed in section 2, the calculation of the normalization term for a log-linear language model is not required for discriminative training criteria including MBR. However, one issue of estimating log-linear weights is the first condition the EBW algorithm requires, i.e., having non-negative coefficients and variables, is no longer valid, because the weights are applied directly to log-likelihood scores. Therefore the EBW re-estimation formulae in equation 11 may be not be directly used to estimate log-linear weights.

To handle this issue in MBR adaptation, the approach adopted in this paper is to normalize the language model scores at the sentence level, by the minimum sentence probability among all recognition hypotheses assigned by all component LMs. This is given by,

$$\log \check{P}(\mathcal{W}) \quad = \quad \log P(\mathcal{W}) - \min_{m,\mathcal{W}} \{\log P_m(\mathcal{W})\} \tag{14}$$

where $\check{P}(\mathcal{W})$ is the normalized LM score for each recognition hypothesis $\mathcal{W}$. First, this will ensure all coefficients and variables in MBR criterion are non-negative and the conditions required by the EBW algorithm valid. Second, because for each sentence all hypotheses' LM scores are normalized by the same term, the posterior distribution over each hypothesis, $P(\mathcal{W}|\mathcal{O})$, remains the same, therefore also the overall MBR criterion in equation 9.

Now the EBW algorithm in equation 11 can be used to estimate the log-linear interpolation weights. The first term of the partial derivative given in equation 12 remains the same as in equation 13. The second term, following the log-linear interpolation given in equation 2, may be derived as,

$$\frac{\partial \log p(\mathcal{O},\mathcal{W})}{\partial \lambda_m} \quad = \quad \sum_i \log \check{P}_m(w_i|h_{i-N+1}^{i-1}) \tag{15}$$

Again, as discussed in section 2 the above derivations may also be applied to for SMT LM adaptation.

4 Implementation Issues

In this section a number of implementation issues that may affect performance of MBR adapted language models are discussed.

Supervision: Like any discriminative self-adaptation scheme, the quality of the initial hypothesis can affect performance of the MBR adapted LM both in recognition and translation. As discussed in section 2, lattice based adaptation provides an natural and elegant extension to 1-best based approach when the quality of the initial hypothesis deteriorates. In the following experiment sections of this paper, N-best based LM adaptation will also serve as a baseline. In order to get the performance upper bound of the adapted models, perplexity

and MBR based adaptation in supervised mode will also be investigated using the correct audio transcription for ASR systems. However, such a comparison is impossible for adapting SMT LMs. This is because the correct English translation based on manual audio segmentation can not be simply "projected" onto the automatic audio segmentation used by the ASR system, due to re-ordering of words and phrases during human translation.

Use of N-best Lists: Multiple hypotheses are required to accumulated the sufficient statistics given in equation 12 for MBR adaptation. This is also true with lattice or N-best based adaptation. In this paper, for both ASR and SMT systems, the top N-best 1000 hypotheses are generated for each speech segment, and kept fixed during language model adaption.

Computation Cost: In order to further reduce the memory requirement, the word probabilities required by the statistics given in equations 5 and 15 are generated off-line for each N-best candidate using each component LM and kept fixed.

Cost Function: MBR adaptation requires the calculation of a cost function in order to improve the task specific evaluation metric being considered. When recognizing Mandarin speech, the cost function used for MBR adaptation is the character error rate. The 1-best output from the ASR system will be used to compute the CER scores of each other entry of the N-best hypotheses generated for every speech segment. For supervised adaptation of ASR LMs, human generated transcription based on manual segmentation must be first mapped onto the automatic audio segmentation used by the ASR system for CER computation. As discussed above, during translation the TER scores will be computed only in unsupervised mode, using the 1-best hypothesis from the SMT system as the reference against all other N-best candidates for each Chinese sentence.

Smoothing Constant D: As discussed in section 3, the setting of the smoothing constant, D, may affect both the optimization stability and generalization. As in standard discriminative training, its setting is largely based on heuristics and empirical results [5]. The form considered in this paper is $D = E \times N_{\mathcal{W}}$, where $N_{\mathcal{W}}$ is the number of Mandarin speech segments to be recognized, or translated, and $E > 0$, typically set as 50. In practice this was found a good compromise between convergence speed and generalization. Varying E was also found having minimum effect on recognition and translation performance. Hence in this paper E is always set as 50 and never altered.

Weights Initialization: This is another factor that may affect the translation performance of MBR interpolated language models. Both equal and PP based weight estimates can be used. The effect of different initialization schemes will be further investigated in section 5.

5 Experiments and Results

In this section experimental results on a Mandarin Chinese broadcast speech transcription and translation task are presented. In the first part, LM adaptation schemes are evaluated on an state-of-the-art Mandarin ASR system. In the second part, machine translation performance using various adapted LMs for the ASR system's output are presented.

5.1 LM adaptation for ASR

The CUHTK Mandarin ASR system was used to evaluate various LM adaptation techniques. The overall structure of the system was similar to that described in [12]. It comprises an initial lattice generation stage using a baseline 58k word list based interpolated 4-gram word language model, and adapted MPE acoustic models trained on HLDA projected PLP features with CMN normalization further augmented with pitch parameters. A total of 942 hours of broadcast news (BN) and broadcast conversation (BC) speech audio data were used for acoustic model training. After text normalization and character to word segmentation, a total of 1.3G words from 20 text sources were used to train an interpolated 4-gram Chinese language model. In the LM adaptation experiments of this paper, only the top 10 Chinese sources with respect to interpolation weights are used to build an interpolated 4-gram Katz style back-off model for lattice rescoring. A generic English language model was also used to handle foreign speech [12]. Information of component LMs and Chinese text sources are give in table 1:

Three Mandarin ASR evaluation sets are used:

Comp LM	Model Size(M)			Text (M)
	2g	3g	4g	
Phoenix	11.50	40.07	8.34	76.89
BC-M	1.19	3.06	3.78	4.83
GIGA2 xin	19.25	26.08	10.39	277.6
BN-M	1.07	2.45	2.91	3.78
GIGA2 cna	24.89	37.05	12.21	496.7
VOARFABBC	2.99	9.24	1.97	30.28
CCTVCNR	5.16	15.23	2.74	26.81
PapersJing	9.43	10.20	11.34	83.73
TDT4	0.71	1.35	0.09	1.76
NTDTV	2.27	1.27	1.23	12.49

Table 1: Model size and text source for Mandarin component LMs.

- bnmdev06: 14 shows, 3.4 hours of BN data broadcast between February 2001 and October 2005 subsuming the RT03 and RT04f Mandarin evaluation data.
- bcmdev05: 5 shows, 2.5 hours of Mandarin BC data broadcast in March 2005.
- eval06: 29 audio snippets, 1.8 hours of Mandarin BN and BC data of the GALE 2006 evaluation set.

Language model adaption schemes were investigated at the audio show level. The form of smoothing constant D described in section 4 was used. A total of 8 iterations of weights re-estimation were performed for MBR adapted LMs. The 1-best output generated by an unadapted, fixed weights interpolated baseline model was used as the supervision for perplexity and MBR adaptation. The top 1000 hypotheses were extracted as the supervision for N-best based adaptation. Component models were finally re-interpolated using the adapted weights to build a back-off 4-gram model for lattice rescoring. Due to the reason discussed in section 3, only linear interpolation based MBR adaptation is considered.

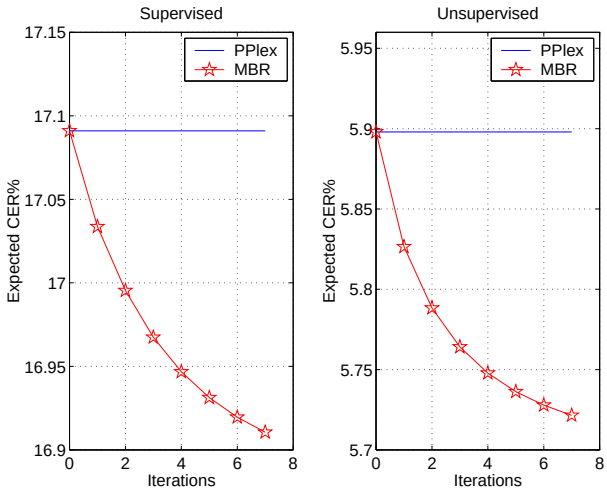

Figure 1: MBR criterion on bnmdev06, bcmdev05 and eval06 for supervised and unsupervised adapted LMs using PP and MBR.

The average expected CER on all three sets for MBR adapted LMs in supervised mode at different iterations in supervised and unsupervised mode are shown in figure 1. The EBW optimization was found fairly stable for the MBR criterion. A steady reduction of expected character error rate can be found against the baseline perplexity adapted model, the starting point of the MBR adaptation. In both cases, approximately 0.2% improvement of MBR criterion were obtained. As expected for unsupervised MBR adaptation the expected error rate is substantially lower.

As discussed in section 4, the initialization of weights may affect the performance of MBR adapted language models. CER performance comparison between using perplexity based, or equal weights initialization is

Sys	Init	CER%		
		bnmdev06	bcmdev05	eval06
fg	pp	8.1	18.8	18.8
	eql	8.1	18.8	18.7
fg-cn	pp	8.0	18.6	18.5
	eql	8.0	18.6	18.5

Table 2: CER performance on bnmdev06, bcmdev05 and eval06 for MBR adaptation using PP or equal weights initialization.

shown in table 2 for all three evaluation sets at both lattice rescoring and the following confusion network (CN) decoding stages. The effect of using different initializations is found small. In the rest of the section, perplexity based interpolation weights are used as the initialization for N-best and MBR adapted models.

Sys	Adapt	CER%		
		bnmdev6	bcmdev05	eval06
fg	fixed	8.4	19.0	19.1
	pp	8.1	18.8	18.9
	nbest	8.1	18.8	18.8
	mbr	8.1	18.8	18.8
fg-cn	fixed	8.3	18.8	18.7
	pp	8.1	18.6	18.5
	nbest	8.1	18.6	18.5
	mbr	8.0	18.6	18.5

Table 3: CER performance of adapted LMs on bnmdev06, bcmdev05 and eval06 for lattice scoring and CN decoding.

CER performance of various adapted LMs are shown in table 3. Absolute CER reductions of 0.3% on bnmdev06, 0.2% on bcmdev05 and 0.3% on eval06 were obtained at the 4-gram lattice rescoring stage using either N-best, or MBR adaptation. Some gains were still retained after CN. The discriminatively adapted MBR model yielded the overall best performance. This can be further illustrated by the crude correlation between word level perplexity and CER scores on this task. Word level perplexity scores for each audio show's 1-best output in bnmdev06 and bcmdev05, selected by the unadapted baseline 4-gram model, are plotted against the show level CER scores in figure 2. This indicates a cost function mismatch when using word level perplexity based LM interpolation for Mandarin ASR.

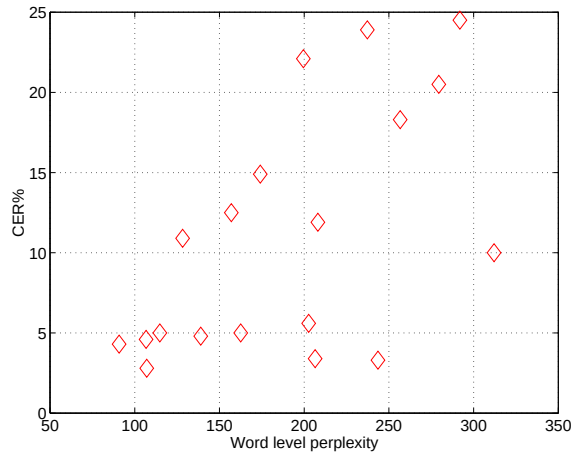

Figure 2: Correlation between word level perplexity and CER

As discussed in section 4, MBR based LM adaptation may be sensitive to the quality of supervision. Hence, it is interesting to obtain an upper bound on performance improvement from MBR adaptation. In table 4, the

4-gram CN stage CER performance of perplexity and MBR based supervised adaptation using reference transcriptions are presented. In order to obtain the CER cost function for MBR adaptation, the human generated manual audio transcriptions were first mapped to the automatic speech segmentation used in the ASR system. As is shown in the table, on this setup MBR based LM adaptation was found insensitive to the supervision error rate.

Adapt	Sup	CER%		
		bnmdev06	bcmdev05	eval06
pp	fg	8.1	18.6	18.5
	ref	8.1	18.6	18.4
mbr	fg	8.0	18.6	18.5
	ref	8.0	18.6	18.4

Table 4: Supervised and unsupervised adapted CER performance on bnmdev06, bcmdev05 and eval06 for PP and MBR adaptation.

Unfortunately the MBR criterion improvement in figure 1 has not been completely projected onto CER reduction in tables 3 and 4 against the perplexity adapted baseline model. This may be because during MBR adaptation rather than the posterior of the best hypothesis with the lowest CER is increased, those of a cluster of other hypotheses with slightly sub-optimal error rates were boosted. This can still lead to an decrease of the expected CER score.

5.2 LM adaptation for SMT

Finally, LM adaptation performance for a SMT system is evaluated. The final output of the above ASR system is post-processed, so that it consists of sentence-like segments via a sentence end detection scheme, and then translated into English text. The MTTK-TTM phrase based translation system was used. Phrase pairs were extracted from word alignments obtained by MTTK on a bilingual parallel Chinese to English corpus consisting of approximately 10 million sentence pairs (220M words on the Chinese side). A weighted finite state transducer based decoding strategy described in [10] was used. Component transducers include a word to phrase segmentation model, phrase reordering model and phrase translation model. A 417k word list based interpolated 4-gram English language model was used to generate the top 1000 hypotheses for later rescoring using various adapted language models. Information of component LMs are give in table 5:

Comp LM	Model Size(M)			Text (M)
	2g	3g	4g	
GIGA2 xin	9.82	13.25	20.21	242.6
BBN	31.50	64.01	110.02	1299.9
MTA	4.73	7.56	12.23	137.6
GIGA2 afp	13.16	25.28	44.47	409.4
GIGA2 apw	20.36	51.61	97.35	921.7
WebNews	2.96	3.43	4.34	44.86
bitex C-E	7.98	12.11	19.17	223.8
CNN	7.24	12.73	20.48	224.2

Table 5: Model size and text source for English component LMs.

Three Mandarin speech translation sets are used, including eval06 as used in previous ASR experiments, and two subsets:

- bnmd06: 7 shows, 1.7 hours pf BN data of bnmdev06.
- bcmd05: 2 shows, 1.2 hours of BC data of bcmdev05.

The remaining BN and BC data of bnmdev06 and bcmdev05 were used to tune the SMT system and therefore not used to evaluate translation performance.

Consistent with the previous experiments for ASR, language model adaption schemes are investigated at the audio show level. Again, the form of smoothing constant D described in section 4 was used. A total of 4 iterations of weights re-estimation were performed for MBR adapted LMs. The 1-best output generated using a unadapted, fixed weights interpolated baseline model was used as the supervision for perplexity and MBR adaptation. Up to 1000 hypotheses were extracted as the supervision for N-best and MBR based adaptation.

Adapt	Int	Init	TER%		
			bnmd06	bcmd05	eval06
fixed	lin	-	72.24	75.28	80.46
pp	lin	eql	72.20	75.26	80.35
nbest	lin	pp	72.21	75.25	80.37
		eql	72.22	75.31	80.52
mbr	lin	pp	72.14	75.23	80.37
		eql	72.16	75.30	80.40
	log	pp	71.73	74.88	79.89
		eql	71.66	74.94	79.84

Table 6: TER performance of adapted LMs on bnmd06, bcmd05 and eval06 for 1000 N-best rescoring.

TER performance of various adapted English language models are shown in table 6 for bnmd06, bcmd05 and eval06. The baseline fixed weights based system gave a translation edit rate of 72.24% on bnmd06, 75.28% on bcmd05 and 80.46% for eval06. Using perplexity based weights adaptation, the TER scores were slightly improved on all sets. Using N-best based adaptation, similar performance were obtained with either perplexity or PP based weights initialization. TER performance of MBR adapted LMs are shown in the final section of the table. Both linear and log-linear interpolation are considered. The linear interpolated MBR model using perplexity based weights initialization marginally outperformed both standard perplexity and N-best based adaptation on the two development sets. The best TER performance were obtained using the log-linear interpolated MBR models. Compared with perplexity based adaptation, the TER scores were improved by 0.47%-0.54% on bnmd06, 0.32%-0.38% on bcmd05 and 0.46%-0.51% on eval06. It is interesting that weights assigned by MBR adaptation are often very different from the perplexity based ones. For example, the TER score of audio show CCTV4_DAILYNEWS_CMN_20060207_145800_12 was improved by 1.78% absolute from MBR adaptation against the perplexity baseline. Using PP based adaptation the top 4 heavily weighted sources are: GIGA2 xin 0.50, bitex C-E 0.31, GIGA2 apw 0.11, BBN 0.06, whilst the PP initialized log-linear MBR adapted model: GIGA2 xin 0.36, BBN 0.28, bitex C-E 0.17, GIGA2 apw 0.14. A similar trend was found on show NTDTV_NTDNEWS12_CMN_20060207_115801_22. Its TER score was reduced by 1.37% absolute from MBR against the perplexity baseline. A substantially higher weight of 0.41 was given to the component LM trained on the BBN text source, in contrast to a much smaller 0.17 determined using perplexity. These suggest MBR adaptation is very different from standard techniques.

6 Conclusion

Unsupervised test-time discriminative adaptation of mixture language models was investigated in this paper for a Mandarin broadcast speech transcription and translation task. A minimum Bayes risk based method is proposed to provide a flexible framework for unsupervised LM adaptation. It generalizes to a variety of forms of recognition and translation error cost functions. An efficient weights re-estimation algorithm was presented for both linear and log-linear interpolated mixture language models. Initial experiments indicate that the correlation between perplexity and character error rate metrics is fairly weak for current Mandarin ASR systems. Performance improvements obtained in both the recognition and translation stages also suggest the proposed form of discriminative LM adaptation may be useful for speech recognition machine translation. Future research will examine integrated discriminative adaptation of translation and language models as a single log-linear model for SMT systems.

References

[1] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts, 1997.

[2] K. Papineni, S. Roukos, T. Ward & W. Zhu, BLEU: a method for automatic evaluation of machine translation, *T.R. RC22176 (W0109-022)*, IBM Research Division, 2001.

[3] M. Snover, B. Dorr, R. Schwartz, L. Micciulla & J. Makhoul, A study of translation edit rate with targeted human annotation, in *Proc. AMTA'06*.

[4] J. Kaiser, B. Horvat & Z. Kacic, A Novel Loss Function for the Overall Risk-criterion Based Discriminative Training of HMM Models, *Proc. ICSLP'00*, Beijing.

[5] D. Povey & P. C. Woodland (2002). Minimum Phone Error and I-smoothing for Improved Discriminative Training, *Proc. ICASSP'02*, Florida, USA.

[6] V. Doumpiotis & W. Byrne. Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition. In *Speech Communication*, (2):142-160, 2005.

[7] V. Goel & W. Byrne. Minimum Bayes-risk automatic speech recognition. In W. Chou & B.-H. Juang, *Pattern Recognition in Speech and Language Processing*. CRC Press, 2003.

[8] F. J. Och & H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. ACL02')*, pp. 295-302, Philadelphia.

[9] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas, & D. Nahamoo (1991). An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems, *IEEE Transactions on Information Theory*, January, 1991.

[10] S. Kumar, Y. Deng & W. J. Byrne. A weighted finite state transducer translation template model for statistical machine translation. *Journal of Natural Language Engineering*, March 2006.

[11] M. J .F. Gales, D. Y. Kim, P. C. Woodland, D. Mrva, R. Sinha & S. E. Tranter. Progress in the CU-HTK broadcast news transcription system, *IEEE Transactions Speech and Audio Processing*, September 2006.

[12] R. Sinha, M. J. F. Gales, D. Y. Kim, X. A. Liu, K. C.Sim, and P. C. Woodland (2006). The CU-HTK Mandarin broadcast news transcription system, *Proc. ICASSP'06*.

[13] M. Tomalin, M.J.F. Gales, X. A. Liu, K.C. Sim, R. Sinha, L. Wang, P.C. Woodland & K. Yu (2007). Improving Speech Transcription For Mandarin English Translation, in *Proc. ICASSP'07*.

[14] D. Mrva & P. C. Woodland (2006). Unsupervised Language Model Adaptation for Mandarin Broadcast Conversation Transcription, in *Proc. ICSLP'06*.

[15] K. Yu & M. J. F. Gales (2007). Bayesian adaptive inference and adaptive training, to appear in *IEEE Transactions Speech and Audio Processing*.

[16] I. Bulyko, S. Matsoukas, R. Schwartz, L. Nguyen & J. Makhoul (2007). Language Model Adaptation in Machine Translation from Speech, in *Proc. ICASSP'07*.

[17] A. Messaoudi, J.-L. Gauvain, & L. Lamel (2006). Arabic transcription using a one million word vocalized vocabulary, in *Proc. ICASSP'06*.

[18] B. Roark, M. Saraclar & M. Collins (2006). Discriminative n-gram language modeling, *Computer Speech and Language*, 2006.

[19] Z. Chen, M. Li & K.-F. Lee (2000). Discriminative Training of Language Model, in *Proc. ICSLP'00*.

[20] Hong-Kwang Jeff Kuo, Eric Fosler-Lussier, Hui Jiang & Chin-Hui Lee (2002). Discriminative Training of Language Models for Speech Recognition, in *Proc. ICASSP'02*.

[21] Hong-Kwang Jeff Kuo & Brian Kingsbury (2007). Discriminative Training of Decoding Graphs for Large Vocabulary Continuous Speech Recognition, in *Proc. ICASSP'07*.

[22] S.-S. Lin & F. Yvon (2005). Discriminative training of finite state decoding graphs, in *Proc Interspeech'05*.