Context Dependent Language Model Adaptation

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

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {xl207,mjfg,pcw}@eng.cam.ac.uk

Abstract

Language models (LMs) are often constructed by building multiple component LMs that are combined using interpolation weights. By tuning these interpolation weights, using either perplexity or discriminative approaches, it is possible to adapt LMs to a particular task. In this work, improved LM adaptation is achieved by introducing context dependent interpolation weights. An important part of this new approach is obtaining robust estimation. Two schemes for this are described. The first is based on MAP estimation, where either global interpolation weights are used as priors, or context dependent interpolation priors obtained from the training data. The second scheme uses class based contexts to determine the interpolation weights. Both schemes are evaluated using unsupervised LM adaptation on a Mandarin broadcast transcription task. Consistent gains in perplexity using context dependent, rather than global, weights are observed as well as reductions in character error rate.

1. Introduction

Back-off n-gram models remain the dominant language modeling approach for state-of-art ASR systems [4]. Training text corpora are often collected from different sources, having a range of topics and styles. One common way of using these multiple sources is to build an n-gram mixture model and tune the interpolation weights (component priors) by minimizing the perplexity on some held-out data similar to the target domain. These weights indicate the "usefulness" of each source for the particular task. To further refine the LM, unsupervised test-set adaptation to a particular broadcast show, for example, may be used. Normally this adaptation scheme only involves updating the interpolation weights as directly adapting n-gram word probabilities is impractical on limited data.

There are two major issues with this standard adaptation scheme. First, each LM training data source will have varying n-gram coverage and associated n-gram frequency depending on the precise nature of the source. Thus the usefulness of a particular source for a domain may vary depending on the word context. When training component LMs for each source, the n-gram "hit-rate" and estimated LM probabilities on the source training data will also be affected by the choice of cut-off setting and smoothing scheme. During LM adaptation, how similar a particular component language model is to the supervision word sequence depends on the contexts being examined. Using global weights take no account of this local variability. Hence, it is preferable to increase the modelling resolution of weight parameters by adding contextual information [3]. Second, the correlation between perplexity and error rate is well known to be fairly weak for current ASR systems. Hence, it may be useful to use discriminative adaptation techniques [6, 8, 2].

To address these issues, this paper investigates the use of discriminatively trained context dependent interpolation weights for unsupervised LM adaptation. As this dramatically increases the number of parameters to estimate, robust weight estimation schemes are required. Two approaches are described in this paper. The first is based on MAP estimation where either the global interpolation weights are used as priors, or context dependent weights estimated on the training data. An important issue when using the training data to estimate interpolation weights is to handle the differences in corpus size. In this work an inverse corpus size weighted version of perplexity, normalized perplexity, is proposed. The second approach uses class history interpolation weights. Rather than using standard approaches to derive the word-to-class mapping, a scheme specifically aimed at class history interpolation weights is proposed. These schemes are evaluated on a state-of-the-art Mandarin broadcast transcription task.

2. Language Model Adaptation

The standard approach for LM adaptation is to adjust the global linear interpolation weights for a mixture model. For word based n-gram models, the log probability of the L word sequence $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$, is given by

$$\ln P(\mathcal{W}) \quad = \quad \sum_{i=1}^{L} \ln P(w_i | h_i^{n-1}) \qquad (1)$$

where w_i denote the ith word of $\mathcal{W}$, and h_i^{n-1} represents its n-gram history of a maximum length of $n-1$ words if available, $< w_{i-1}, w_{i-2}, ..., w_{i-n+1} >$. The interpolated word probability is computed as,

$$P(w_i | h_i^{n-1}) \quad = \quad \sum_m \lambda_m P_m(w_i | h_i^{n-1}) \qquad (2)$$

where λ_m is the global weight for the m^{th} component model.

The interpolation weights are often found based on **perplexity** (PP) measure, $\text{PP} = \exp\{-\ln P(\mathcal{W})/L\}$. However it is also possible to use discriminative approaches, such as minimum Bayes risk (MBR), to estimate the weights [6]. Assuming "sufficient" adaptation data are available, the ML weights may be found using Baum-Welch (BW) estimation with ML context independent statistics, $\mathcal{C}_m^{\mathsf{ML}}(\mathsf{null})$,

$$\mathcal{C}_m^{\mathsf{ML}}(\mathsf{null}) \quad = \quad \tilde{\lambda}_m \left. \frac{\partial \ln P(\mathcal{W})}{\partial \lambda_m} \right|_{\lambda=\tilde{\lambda}} \qquad (3)$$

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

where $\tilde{\lambda}$ is the current weight estimate. For MBR training the extended Baum-Welch algorithm can be used [6]. Here the context independent discriminative statistics are,

$$\mathcal{C}_m^{\mathsf{MBR}}(\mathsf{null}) \;=\; \tilde{\lambda}_m \left.\frac{\partial \mathcal{F}_{\mathsf{MBR}}(\lambda)}{\partial \lambda_m}\right|_{\lambda=\tilde{\lambda}} + D \qquad (4)$$

where D is a regularization constant controlling the convergence speed in the same form as in [6]. The exact forms of the partial derivatives in (3) and (4) are given in [6].

3. Context Dependent LM Adaptation

As discussed, the global weights assigned to component n-gram language models take no account of the surrounding contexts. In order to incorporate more context information, a more general form is a context dependent weight, $\boldsymbol{\phi}(h)$. Using an n-gram context history, the interpolated word probability in (2) thus becomes,

$$P(w_i|h_i^{n-1}) \;=\; \sum_m \boldsymbol{\phi}_m(h_i^{n-1}) P_m(w_i|h_i^{n-1}) \qquad (5)$$

where $\boldsymbol{\phi}_m(h_i^{n-1})$ is the m^{th} component weight vector for n-gram history h_i^{n-1}. The same Markov chain assumption of n-gram models is made such that the interpolation weights for word w_i only depends on the preceding $n-1$ words. These context dependent weights are also constrained to be positive and sum-to-one. They can be ML or discriminatively trained using the BW or EBW algorithm as above, but with context dependent statistics $\mathcal{C}_m(h_i^{n-1})$. However, as the history length grows, the number of interpolation weights to estimate increases exponentially. As often only limited adaptation data is available, robust weight estimation schemes are required.

MAP Weight Adaptation: One approach to address the robustness issue is to use MAP estimation. Take the perplexity based estimation as an example, this is given by

$$\hat{\boldsymbol{\phi}}_m(h_i^{n-1}) \;=\; \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau \boldsymbol{\phi}_m^{\mathsf{Pr}}(h_i^{n-1})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau} \qquad (6)$$

where, $\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1})$ is context dependent version of the global statistics in 3, and τ controls the contribution from an appropriate weight prior, $\boldsymbol{\phi}_m^{\mathsf{Pr}}(h_i^{n-1})$. The same form of smoothing may be used for MBR adaptation. One key issue with MAP adaptation is the choice of smoothing prior. Two forms are considered in this work:

(a) ***dynamic context independent*** prior using global weights estimated on the supervision data during test-time.

(b) ***static context dependent*** prior based on the interpolation weights estimated on the training data.

Using a context dependent prior gives a finer modelling resolution. However, for robust prior estimation this requires the use of the training data. When using multiple text sources, the sufficient statistics derived from equations (1) and (5) in re-estimation will be dominated by large sized corpora, and thus introduce a bias. This is a fundamental issue that can affect the adaptation performance. For a more general model interpolation scheme using the training data, this issue is also present and must be addressed. The approach proposed in this paper is to use a corpus length normalization scheme. The training data log-probability given in equation (5) is modified as,

$$\ln P_{\mathsf{norm}}(\mathcal{W}) \;=\; \sum_q \frac{L}{L_q} \sum_{i=1}^{L_q} \ln P(w_i|h_i^{n-1}) \qquad (7)$$

where L_q is the total number of words in the q^{th} corpus. The **normalized perplexity** (nPP) measure is computed using the above as $\mathrm{nPP} = \exp\{-\ln P_{\mathsf{norm}}(\mathcal{W})/\sum_q L\}$.

Using the nPP criterion, weights are determined by the *average* word probability from each data corpus. The bias to larger sized corpora may be handled. The weights associated with each of the source language models are determined by a number of attributes including n-gram coverage and form of probability estimation. If certain sources are known to be useful for the particular domain of interest, for example the acoustic transcriptions, it is possible to bias the weight estimation during the LM construction. If low cut-offs are used then the source specific language models will have high probabilities on their training data compared to others with higher cut-off settings. Similarly if robust discounting schemes are used then the model will also generalize well on other data. By using this prior knowledge general languages models with context dependent interpolation weights may be estimated on the training data. They can be used as priors in adaptation, or standard LMs in decoding.

For contexts that are not seen in the training or adaptation data, rather than using global priors or weights, a simple back-off strategy can be used

$$\boldsymbol{\phi}_{\mathsf{bo}}(h_i^{n-1}) \;=\; \begin{cases} \boldsymbol{\phi}(h_i^{n-1}) & \text{if } \exists\, \boldsymbol{\phi}(h_i^{n-1}) \\ \boldsymbol{\phi}(h_i^{n-2}) & \text{else if } \exists\, \boldsymbol{\phi}(h_i^{n-2}) \\ \cdots & \cdots \\ \boldsymbol{\phi}(\mathsf{null}) & \text{otherwise} \end{cases} \qquad (8)$$

Class Context Dependent Weights: Class-based n-gram models have been shown to be helpful in addressing the data sparsity problem [1]. Words are clustered into syntactically, semantically or statistically equivalent classes. The intuition is that even if a word n-gram does not occur in the training data, the corresponding class n-gram may be available. To more robustly handle unseen or rarely observed events, a class based approach may also be used for context dependent weight estimation. Due to the high dimensionality of the history space, this is generally non-trivial for longer range contexts. The approach considered here is to perform the clustering at the word level. Interpolation weights are then shared among word histories that can be mapped to the same sequence of classes. This is given by,

$$P(w_i|h_i^{n-1}) \;=\; \sum_m \boldsymbol{\phi}_m(\mathcal{G}_i^{n-1}) P_m(w_i|h_i^{n-1}) \qquad (9)$$

where $\mathcal{G}_i^{n-1}$ is the preceding $n-1$ class history determined by a unique word to class mapping. Handling of rare and unseen class contexts uses the same smoothing and back-off schemes as word based weights.

One key issue is how to derive a suitable word to class mapping. An efficient clustering scheme, referred to as *exchange algorithm*, has been proposed and widely used for standard class based n-gram models [5]. However, this algorithm may not be appropriate for context dependent weights. Therefore, an alternative clustering algorithms is required. The method considered is a maximum likelihood based *weight merging* algorithm explicitly derived for context dependent weights.

```
initialize each word in vocabulary as a distinct class;
for each word as history train 1-gram weight using ML;
iterate until the target number of classes obtained:
  find the pair of word classes whose merging gives the
  maximum likelihood gain or minimum loss;
  merge the found class pair into one class.
```

As discussed in section 3, in order to reduce the bias to larger corpora in the clustering data, the normalized log-likelihood of equation (7) will be used. Note that directly computing the likelihood using equation (9) is infeasible, due to the iterative nature of weight estimation and the memory requirement to store component n-gram probabilities for all words in the clustering data. To handle this problem, the log-likelihood lower bound derived using Jensen's inequality is used instead:

$$\ln P_{\text{norm}}(\mathcal{W}) \geq \sum_q \frac{L}{L_q} \sum_{i,m} \phi_m(g(w_{i-1})) \ln P_m(w_i|h_i^{n-1}) \quad (10)$$

Thus the sufficient statistics for likelihood computation simplify to the sum of component n-gram log probabilities for each word, w, that have itself as the immediate proceeding history, $\left\{\sum_{i,h_i^1=w} \ln P_m(w_i|h_i^{n-1})\right\}$. The change of log-likelihood bound only depends on the current pair of classes being merged, while all other classes are fixed. The weight estimates after a merging step can be derived from combining the ML weight statistics of the two classes before the merge, as given in (3).

4. Experiments and Results

The CU-HTK Mandarin ASR system was used to evaluate various LM adaptation techniques [9]. It comprises an initial lattice generation stage using a 58k word list, interpolated 4-gram word based back-off LM, and adapted MPE acoustic models trained on 942 hours of broadcast speech data. A total of 1.0G words from 10 text sources were used in LM training. Information on corpus size, cut-off settings and smoothing schemes for component LMs are give in table 1. In order to reduce the bias to large corpora as discussed in section 3, minimum cut-offs and modified KN smoothing were used for smaller sources, for example, the two acoustic transcription sources, bcm and bnm. For the two largest corpora, giga-xin and giga-cna, more aggressive cut-offs and Good Turing (GT) discounting were used.

Comp LM	Text (M)	Train Config	Global Weight Tuning		
			PPTest	PPTrn	nPPTrn
bcm	4.83	111,kn	0.2325	0.0049	0.1426
bnm	3.78	111,kn	0.1501	0.0066	0.1729
giga-xin	277.6	123,gt	0.1036	0.2428	0.1079
giga-cna	496.7	123,gt	0.0791	0.4577	0.0815
phoenix	76.89	112,kn	0.1542	0.1125	0.1225
voarfabbc	30.28	112,kn	0.0966	0.0270	0.0734
cctvcnr	26.81	112,kn	0.0592	0.0391	0.0844
tdt4	1.76	112,kn	0.0318	0.0060	0.0717
papersjing	83.73	122,kn	0.0444	0.0919	0.0928
ntdtv	12.49	122,kn	0.0485	0.0114	0.0503

Table 1: Text source, 2/3/4-gram cut-off settings, smoothing scheme used in training and global ML weights tuned using test set PP, training data PP and nPP scores for component LMs.

Three Mandarin broadcast speech evaluation sets were used: bndev06 of 3.4 hour BN data, bcdev05 of 2.5 hours of BC data and the 1.8 hour GALE 2006 evaluation set eval06 [6]. An additional held-out set dev07 of 30k words was used for PP evaluation. PP and nPP scores for various globally interpolated LMs are presented in table 2. The first two lines of the table show the performance of using equal, or global weights tuned on test data PP scores. The latter is the standard form of model interpolation for current ASR systems. These weights are in the fourth column of table 1. The third line shows the performance of weights tuned using the training data PP metric. Large PP reductions were obtained on the training data. However, there was a significant degradation of 77 points on the test data PP score. This is due to the corpus size bias discussed in section 3. Such a bias further manifests itself in the global weights given in the 5th column of table 1. The largest two corpora, giga-cna (0.46) and giga-xin (0.24) were heavily weighted.

Wgt Est	Trn Crit	PP Trn	nPP Trn	PP Test	PP dev07
-	Eql	169.0	84.1	227.3	239.7
Test	PP	184.0	84.0	214.1	229.5
Trn	PP	117.5	130.8	304.3	296.3
	nPP	176.6	82.0	219.6	228.6

Table 2: PP and nPP scores of globally weighted LMs on training and test data (combined bndev06+bcdev05+eval06 set) and a held-out set dev07.

Using the nPP metric in weight estimation, this bias was greatly reduced, as given in the last line of table 2. The corresponding weights are in the 6th column of table 1. Large sized corpora no longer dominate the weight assignment. As discussed in section 3, weights are determined by a combination of fitness to observed data and generalization to other sources. For example, the biggest giga-cna source, of Taiwanese origin and different in style from other sources, trained using aggressive cut-offs and simple GT discounting, is now weighted by 0.082. A PP reduction of 7.7 points (3.4% rel) was obtained on the test data against the equal weight baseline. It is interesting that the nPP tuned weights are close to those tuned on the other three sets, and also gave the best PP on the held-out set dev07. These results suggest the nPP criterion may be used as an alternative LM interpolation technique. For robust estimation of context dependent interpolation weights on the training data, the nPP based approach becomes even more useful.

Hist Type	Clustering Algorithm	Num Class	PP Trn	nPP Trn	PP Test
word	-	-	150.3	69.7	209.8
class	exchange algorithm	50	166.6	76.1	217.9
		100	164.4	74.8	216.6
		200	160.9	73.5	214.8
		400	159.7	73.0	213.9
	weight merge	50	165.0	73.6	213.5
		100	163.5	73.4	213.1
		200	162.4	73.1	213.0
		400	161.2	72.8	212.7

Table 3: PP and nPP scores of interpolated LMs on training and test data using nPP trained 1-gram history dependent weights.

Compared with the nPP tuned, and globally interpolated system in table 2, further test set PP gains of 10 points were obtained using 1-gram history dependent weights, as shown in the first line of table 3. This model also outperformed the test data PP tuned baseline in table 2 by 4.2 points on test data PP, and the equally weighted model by 17.5 points (7.7% rel). The remaining part of table 3 shows the performance of various models using class dependent weights with different numbers of clusters. In all configurations, the bottom-up merging algorithm proposed in section 3 outperformed the baseline exchange algorithm on both training data nPP and test data PP scores, though not against using word based weights. This is expected as the merging algorithm evaluates a more appropriate log-likelihood function during clustering for context dependent weights. In both tables 2 and 3 there is a consistent and strong correlation between training data nPP and test set PP scores.

Now it's interesting to examine the performance of adapted LMs using context dependent interpolation weights. LM adaptation was performed at the audio show level. Unless otherwise stated, equal weight initialization was used. Settings of the global dynamic smoothing prior $\tau = 10$ and constant D given in section 3 were used. A total of 8 iterations of weights re-estimation were performed. The top 1000 hypotheses were extracted for MBR adaptation. Component models were finally re-interpolated using adapted weights to build a back-off 4-gram model for lattice rescoring. The 4-gram lattice 1-best output generated by a standard globally tuned baseline (second line of table 2) was used as the adaptation supervision. The 100 classes derived using the weight merging algorithm given in the bottom section of table 3 were used. Performance of ML adaptation are shown in table 4.

Adapt (Prior)	Hist Len	Hist Type	CER%/PP(Ref.)		
			bndev06	bcdev05	eval06
-	-	-	8.4/194	19.0/265	19.1/325
	1g	word	8.3/190	19.0/266	19.2/319
global (-)	-	-	8.1/161	18.8/238	18.9/288
context (a)	1g	word	8.1/148	18.8/222	18.8/269
		c100	8.1/157	18.8/232	18.7/280
	3g	word	8.1/147	18.8/221	18.8/269
		c100	8.1/151	18.8/226	18.7/272
context (b)	1g	word	8.1/130	18.9/212	18.9/240

Table 4: CER and PP performance of ML adapted 4-gram LMs on bndev06, bcdev05 and eval06 for lattice scoring.

Using global PP adaptation, there are 27 to 37 points of PP improvements (10% to 17% rel) for all sets over the unadapted baseline system shown in the first line of table 4. Absolute CER gains of 0.3% on bndev06 and 0.2% on the other two sets were obtained. Using context dependent adaptation, a further PP reduction of 13 to 18 points (8% rel) was obtained by the word level 1-gram weights. However, the CER gains were marginal. The more compact 100 class based system gave better CER performance, but higher PP scores for all sets. Using longer 3-gram weights gave varying PP improvements, but no CER gains. As discussed in section 3, it is also interesting to use the nPP metric estimated static weights shown in the second line of table 4 as a prior for context dependent adaptation. This system is shown in the last line of table 4. Despite further PP reductions of 10 to 29 points on all sets, and 19% relative on bndev06 over the "global" baseline, no CER gain was obtained. These trends may be due to the weak correlation between PP and error rate. Adaptation may improve PP on the common contexts observed in both the supervision and reference, but not necessarily helpful in generalization and discrimination. Hence, it would be interesting to evaluate the performance of MBR adaptation.

These are shown in table 5. For 1-gram weights the MBR systems gave comparable performance to the ML baselines in table 4. Increasing the history length to 3 further reduced the CER. The best performance was obtained using 3-gram word history based weights, which gave absolute CER gains of 0.3% on bndev06, 0.4% on bcdev05 and 0.5% on eval06 over the unadapted baseline (statistically significant).

Adapt (Prior)	Hist Len	Hist Type	CER%		
			bndev06	bcdev05	eval06
-	-	-	8.4	19.0	19.1
mbr (a)	1g	word	8.1	18.8	18.8
		c100	8.1	18.8	18.7
	3g	word	***8.1***	***18.6***	***18.6***
		c100	8.1	18.7	18.7

Table 5: CER performance of discriminatively adapted 4-gram LMs on bndev06, bcdev05 and eval06 for lattice scoring.

5. Conclusion

Unsupervised test-time context dependent adaptation of n-gram mixture models using a discriminative method was investigated in this paper. Map-based adaptation of back-off weights and class based schemes were used to address the data sparsity problem. Two forms of smoothing priors were proposed. An efficient bottom-up maximum likelihood clustering algorithm was derived for the class based approach. Initial experiments on a state-of-the-art Mandarin broadcast speech transcription task suggest that the proposed technique may be useful for speech recognition. Future research will focus on integrated discriminative weight clustering and estimation, and a hierarchical interpolation in model training and adaptation.

6. References

[1] P. F. Brown et al. (1992). Class-based n-gram models of natural language. *Computational Linguistics* 18 (4).

[2] I. Bulyko, S. Matsoukas et al. (2007). Language Model Adaptation in Machine Translation from Speech, in *Proc. ICASSP'07*.

[3] B. Hsu (2007), Generalized Linear Interpolation of language Models (2007). *Proc. IEEE ASRU'07*, Kyoto.

[4] S. M. Katz (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. ASSP* 35 (3), 400401.

[5] R. Kneser and H. Ney (1993), "Improved clustering techniques for class based statistical language modeling," in *Proc. of Eurospeech93'*, Berlin.

[6] X. Liu, W. J. Byrne, M. J. F. Gales, P. C. Woodland et al. (2007). Discriminative Language Model Adaptation for Mandarin Broadcast Speech Transcription and Translation, in *Proc. IEEE ASRU'07*, Kyoto.

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

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

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