Use of Contexts in Language Model Interpolation and 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 component models on multiple text sources to be combined using global, context free interpolation weights. By re-adjusting these weights, LMs may be adapted to a target domain representing a particular genre, epoch or other higher level attributes. A major limitation with this approach is other factors that determine the "usefulness" of sources on a context dependent basis, such as modeling resolution, generalization, topics and styles, are poorly modeled. To overcome this problem, this paper investigates a context dependent form of LM interpolation and test-time adaptation. Depending on the context, a discrete *history weighting function* is used to dynamically adjust the contribution from component models. In previous research, it was used primarily for LM adaptation. In this paper, a range of schemes to combine context dependent weights obtained from training and test data to improve LM adaptation are proposed. Consistent perplexity and error rate gains of 6% relative were obtained on a state-of-the-art broadcast recognition task.

1. Introduction

In ASR systems language models (LMs) are often constructed by training n-gram components models [3] on data from a set of diverse sources. These are then combined using a global level weighted probability interpolation. To reduce the mismatch between the interpolated model and target domain, these global, context free interpolation weights may be tuned by minimizing the perplexity on some held-out data. They indicate the "usefulness" of each source for a particular task. To further improve robustness to varying styles or tasks, unsupervised test-time adaptation to a particular broadcast show, for example, may be used. As directly adapting n-gram probabilities is impractical on limited amounts of data, standard adaptation schemes only involve updating the global interpolation weights.

There are two issues with the above standard methods. First, the diversity among data sources manifests itself in a wide range of factors. The precise nature of each source is jointly determined by a combination of multiple attributes. Some may be sufficiently modeled on a higher level using global, context *independent* weights, for instance, source of collection, epoch and genre. Others factors including n-gram modeling resolution and generalization, topics and styles, affect the contribution of sources on a local, context *dependent* basis. Thus the usefulness of a particular source can vary depending on the word context for both LM interpolation and adaptation. Global weights take no account of such local variability. Hence, it is preferable to increase the modeling resolution of weight parameters by adding contextual information [1, 2, 5]. Second, the correlation between perplexity and error rate is known to be fairly weak for current ASR systems. Hence, it may be useful to use discriminative training techniques [6, 4].

To address these issues, this paper investigates the use of context dependent interpolation in both training and test-time self-adaptation of language models. Under this framework, minimum Bayes risk (MBR) based discriminative training schemes are also investigated. To handle the data sparsity issue, several robust estimation schemes for context dependent weights are first reviewed in section 2. In this paper a range of methods to combine context dependent weights obtained from training and test data to improve LM adaptation are proposed in section 3. Experimental results on a state-of-the-art Mandarin broadcast speech transcription task are presented in section 4.

2. Context Dependent Language Model Interpolation and Adaptation

In standard word based n-gram mixture LMs, the linearly interpolated probability of the i^{th} word of a L word long sequence $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$ is given by

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

where w_i denotes the ith word of $\mathcal{W}$, h_i^{n-1} represents its history of $n-1$ words maximum, $< w_{i-n+1}, ..., w_{i-1} >$, and λ_m is the global, context free weight for the m^{th} component.

As discussed, the above takes no account of surrounding contexts. In order to incorporate more context information, a more general form is to introduce a context dependent *history weighting function*, $\phi(h)$, to dynamically adjust the contribution from component models. Thus equation (1) is extended to

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

where $\phi_m(h_i^{n-1})$ is the m^{th} component weight for history context h_i^{n-1}. By default they are constrained to be positive and sum-to-one. A history weighting function can be in either discrete or continuous form. In this paper only discrete forms are considered. They can be represented by a tree structured hierarchy of context dependent interpolation weights. An example is shown in figure 1 for tri-gram LMs. Such hierarchy will be extensively used in the rest of this paper. As the number of weight parameters to estimate increases exponentially with context length, robust weight estimation schemes are required.

Interpolation Using Training Data: Text data for LM training are often available in large quantities, e.g., in billions of

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.

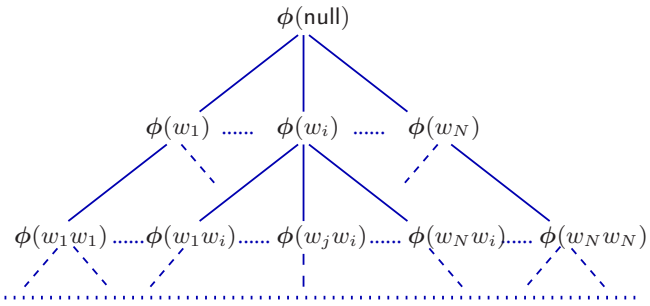

Figure 1: Hierarchy of context dependent interpolation weights for tri-gram back-off LMs with a maximum two word history.

words [4]. Using training data, robust estimation of context dependent weights may be ensured. In this paper, the use of training data is only considered for perplexity (ML) based estimation. For discriminative training, this is non-trivial because confusable word sequences have to be explicitly generated for text genre sources, in addition to audio transcription.

The perplexity (PP) metric $\mathcal{F}_{\text{PP}} = \exp\{-\ln P(\mathcal{W})/L\}$, is computed using the entire word sequence's log-probability $\ln P(\mathcal{W}) = \sum_{i=1}^{L} \ln P(w_i|h_i^{n-1})$. When optimizing PP on training data of multiple text sources, the sufficient statistics for weights estimation will be dominated by large sized corpora, and thus introduce a bias [5]. This bias can be removed using a normalized log-probability,

$$\ln P_{\text{norm}}(\mathcal{W}) = \sum_m \frac{L}{L_m} \sum_{i=1}^{L_m} \ln P(w_i|h_i^{n-1}) \quad (3)$$

where L_m is the total number of words in the m^{th} corpus. The associated *normalized perplexity* (nPP) metric is then computed as $\mathcal{F}_{\text{nPP}} = \exp\left\{-\ln P_{\text{norm}}(\mathcal{W})/\sum_m L\right\}$.

As discussed in section 1, the variability among data sources and their contribution are jointly determined by a combination of multiple attributes. Some of them may be sufficiently modeled using global, context independent weights, for example, epoch and genre. Others such as modeling resolution, topics and styles, require local, context dependent weighting. Using the nPP criterion, interpolation weights at both levels can be estimated. Often the global level diversity among sources is further enlarged by conscious decisions when building component models. If certain sources are known to be useful for the domain of interest, for example, acoustic transcriptions, a bias to components of the same genre may be introduced during LM construction. When low cut-offs are used for these sources, the associated component models will have high probabilities on their training data compared to others built with more punitive cut-off settings. Similarly if robust discounting schemes are used then these models will also generalize well on other data. Using the nPP criterion, general LM interpolation using both context free and dependent weights may be robustly estimated on the training data. They can be used as standard LMs for decoding prior to test-time domain adaptation.

MAP Estimation: One common approach to address the robustness issue is to use maximum *a-posteriori* (MAP) estimation. Take the perplexity or ML based adaptation as an example, this is given by

$$\hat{\phi}_m(h_i^{n-1}) = \frac{\mathcal{C}_m^{\text{ML}}(h_i^{n-1}) + \tau\phi_m^{\text{Pr}}(h_i^{n-1})}{\sum_m \mathcal{C}_m^{\text{ML}}(h_i^{n-1}) + \tau} \quad (4)$$

where $\mathcal{C}_m^{\text{ML}}(h_i^{n-1})$ is context dependent ML statistics for history context h_i^{n-1}, and τ controls the contribution from weight prior, $\phi_m^{\text{Pr}}(h_i^{n-1})$. One key issue with MAP estimation is the choice of smoothing prior. In previous research the global, context free weights, $\phi(\text{null})$, was used [5]. In order to introduce more context information, rather than completely backing off to the context independent weights, a hierarchical smoothing using weights of lower order contexts is used in this paper. Take the perplexity based estimation as an example, this is given by

$$\hat{\phi}_m(h_i^{n-1}) = \frac{\mathcal{C}_m^{\text{ML}}(h_i^{n-1}) + \tau\hat{\phi}_m(h_i^{n-2})}{\sum_m \mathcal{C}_m^{\text{ML}}(h_i^{n-1}) + \tau} \quad (5)$$

The form of hierarchical smoothing in equation (4) can also be applied to nPP and MBR based statistics.

Weight Back-off for Unseen Contexts: For contexts unobserved in the training or adaptation data, a back-off recursion using the tree structure in figure 1 is performed,

$$\phi^{\text{bo}}(h_i^{n-1}) = \begin{cases} \phi(h_i^{n-1}) & \text{if } \exists\, \phi(h_i^{n-1}) \\ \phi(h_i^{n-2}) & \text{else if } \exists\, \phi(h_i^{n-2}) \\ \cdots & \cdots \\ \phi(\text{null}) & \text{otherwise} \end{cases} \quad (6)$$

which may eventually simplify to the global, context free weights, $\phi(\text{null})$. In contrast to standard n-gram models, no normalization term is required.

3. Weight Set Combination

When adapting LMs using context dependent interpolation, two sets of weights are available. These are obtained from the training data nPP estimation and test adaptation respectively:

- ***training*** data nPP weights estimated using a hierarchical smoothing. They provide richer context information and finer modeling resolution, but potentially larger mismatch against the target domain during LM adaptation.
- ***test*** data self-adapted weights using equation (5) and a hierarchical smoothing. These provide a closer match to the target domain of interest. However, as the supervision may contain errors and not all contexts in the reference can have their own weights, a back-off to a lower order context based weights using equation (6) is necessary. This will result in reduced modeling resolution.

The above two sets of weight information provide either domain neutral, longer contexts based weights, or in-domain, shorter contexts based ones. In order to balance this trade-off, it is preferable to appropriately combine the two for context dependent LM adaptation. Previous research largely relied on test set information. The combined use of the above two was very limited [5]. In this section four weight combination schemes are proposed to incorporate both training and test set information. They can be categorized into two broad types of techniques: two-stage MAP estimation and log-linear weight combination. Within each category, it is also optional to further supplement the adapted weights of contexts obtained from the training data with weights of contexts uniquely observed in the test set supervision. These contexts may carry additional information of the target domain for adaptation, and it is thus interesting to include them via a union operation. Note that the hierarchical smoothing of equation (5) effectively uses a lower order context based weight prior. However, for clarity in the rest of this paper the term "prior" is reserved and exclusively refers to nPP weights estimated on the training data.

A. Two-stage MAP Estimation: In the first stage nPP based LM interpolation is performed. Contexts are extracted from the training data. To improve robustness, their weights are MAP estimated using a hierarchical smoothing as in equation (5). In the second stage, test-time LM self-adaptation is performed, where the nPP estimated context dependent weights are used as a prior. For example, for ML based adaptation, the final adapted m^{th} component weight of history context h_i^{n-1} is given by

$$\hat{\phi}_m^{\text{comb}}(h_i^{n-1}) = \frac{\mathcal{C}_m^{\text{ML}}(h_i^{n-1}) + \tau \hat{\phi}_m^{\text{nPP}}(h_i^{n-1})}{\sum_m \mathcal{C}_m^{\text{ML}}(h_i^{n-1}) + \tau} \quad (7)$$

Note that in both stages all contexts are exclusively obtained from the training data.

B. Two-stage MAP Estimation and Union: This is an extension of option **A**. As discussed, contexts uniquely observed in the test set supervision may carry additional useful information of the target domain, their associated weights are also MAP adapted to the supervision using equation (5) and merged into the final combined context dependent weight set, rather than being discarded. The MAP adapted training data weight tree of figure 1 is effectively expanded by adding more nodes that represent the newly observed histories in the supervision.

C. Log-linear Composition: MAP estimation may be viewed as a weighted linear interpolation, for example, between an ML estimate and its smoothing prior. There are two issues with this approach. First, training set contexts that are unavailable in the supervision will back-off to a domain neutral nPP prior containing minimum information of the test data. Second, due to the nature of linear interpolation, test set weights that are MAP adapted to incorrect supervision using option **A** can retain certain mis-ranking of component LMs. To address these issues, an alternative is to use a log-linear composition. During a finite state transducer composition between two back-off context dependent weight sets, the longest matching contexts from each will be automatically found and used via the back-off process given in equation (6), rather than using zero test set information for unseen events as in equation (7). Furthermore, a log-linear interpolation can also reject component LM weighting obtained from an erroneous supervision that are very different from the training set nPP prior. Therefore it can improve robustness of weight combination. In the first step, in common with option **A**, contexts are extracted from the training data and their associated nPP weights MAP estimated using a hierarchical smoothing. In the second step, contexts are extracted from the test set supervision and their weights MAP estimated using a hierarchical smoothing of equation (5). The final combined weights are,

$$\hat{\phi}_m^{\text{comb}}(h_i^{n-1}) = \frac{\hat{\phi}_m^{\text{nPP}}(h_i^{n-1})^{\alpha} \hat{\phi}_m^{\text{bo}}(h_i^{n-1})}{\sum_m \hat{\phi}_m^{\text{nPP}}(h_i^{n-1})^{\alpha} \hat{\phi}_m^{\text{bo}}(h_i^{n-1})} \quad (8)$$

where α is a tunable log-linear scaling factor that controls the contribution from the nPP prior. In this option it is un-tuned and set as $\alpha = 1.0$. In the same fashion as **A**, new contexts unique to the test set supervision are discarded.

D. Weighted Log-linear Composition and Union: This is a modified form of option **C**. For any context extracted from the training data, if it has no matching context of any length at all in the test set supervision and therefore completely backs off to the context free, global weights, equation (8) ($\alpha = 1.0$) is still used to obtained the combined weights. Otherwise, the test data supervision adapted weights for the longest matching context will be used. This is effectively achieved by setting $\alpha = 0$ in equation (8). In common with **B**, weights of contexts uniquely observed in the test set supervision are also added. Compared with **C**, this approach is leaning more to the estimates from the test set supervision whenever context dependent weights are available. Hence, it is closer to the target domain for LM adaptation.

4. Experiments and Results

The CU-HTK Mandarin ASR system was used to evaluate LMs using various interpolation and adaptation techniques [4]. 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, smoothing schemes and component LMs are given in table 1. For data sources that are closer in genre to the test data, minimum cut-offs and modified KN smoothing were used. These include two audio transcriptions sources, bcm and bnm, and additional web data from major TV channels such as CCTV and Phoenix TV. For the two largest corpora of newswire genre, giga-xin and giga-cna, more aggressive cut-offs and Good Turing (GT) discounting were used. Three Mandarin broadcast speech evaluation sets were used: bn06 of 3.4 hour broadcast news (BN) data, bc05 of 2.5 hours of broadcast conversation (BC) data and the 1.8 hour GALE 2006 evaluation set eval06.

Comp LM	Text (M)	Train Config	Model Size(M) 2g	3g	4g
bcm	4.83	111,kn	1.19	3.06	3.78
bnm	3.78	111,kn	1.07	2.45	2.91
giga-xin	277.6	123,gt	19.25	26.08	10.39
giga-cna	496.7	123,gt	24.89	37.05	12.21
phoenix	76.89	112,kn	11.50	40.07	8.34
voarfabbc	30.28	112,kn	2.99	9.24	1.97
cctvcnr	26.81	112,kn	5.16	15.23	2.74
tdt4	1.76	112,kn	0.71	1.35	0.09
papersjing	83.73	122,kn	9.43	10.20	11.34
ntdtv	12.49	122,kn	2.27	1.27	1.23

Table 1: 2/3/4-gram cut-off settings, smoothing scheme used in training, and model size information for text sources.

PP and nPP scores for two interpolated LMs are shown in table 2. The first LM uses global, context free weights that are perplexity tuned on bn06+bc05. It serves as the baseline LM in this paper. The second system uses 3-gram word history based context dependent nPP interpolation. Statistically significant CER improvements of 0.3% and 0.4% absolute were obtained over the baseline LM on bn06 and eval06 respectively.

Intplt Crit	Context	CER%/PP(Reference) bn06	bc05	eval06
base	-	8.4/195	19.0/227	19.1/232
npp	3g	8.1/179	19.0/213	18.7/215

Table 2: PP scores and lattice rescoring 1-best CER% performance of interpolated LMs on bn06, bc05 and eval06.

Performance of ML adapted LMs are shown in table 4. For standard test set adaptation without using training set information, the form of MAP adaptation in equation (5) with a hierarchical smoothing was used. The smoothing constant was set as

$\tau = 2.5$ in the experiments. A total of 8 iterations of weights re-estimation were performed. The first line is the baseline system in table 2 using context free interpolation. The 4-gram lattice 1-best hypothesis it generated was used as the adaptation supervision. Note that it is also possible to use outputs from the nPP system in table 2 as the adaptation supervision. However, this was found to give no performance improvement without being further combined with the nPP prior weight set. For example, when adapting 3-gram context based interpolation weights, simply using the nPP system's outputs as the supervision does not reduce the error rate, as is shown in table 3.

Adapt Supv	Adapt Context	CER%/PP(Reference) bn06	bc05	eval06
base	3g	8.0/133	18.8/176	18.7/184
npp		8.0/132	18.8/176	18.7/181

Table 3: Performance of ML adapted LMs with 3-gram context weights using different supervision on bn06, bc05 and eval06.

Performance of three perplexity adapted systems with context free or dependent weights are shown from the 2nd to 4th line. Using PP adaptation of global weights, 26 to 45 points of PP improvements (12%-23% rel) were obtained for all sets over the unadapted baseline system in the first line of table 4. Absolute CER gains of 0.3% on bn06 and eval06 were obtained. Using context dependent form of PP adaptation, further PP reductions were observed, but the CER gains were marginal.

Context Prior	Context Adapt	Wgt Com	CER%/PP(Reference) bn06	bc05	eval06
-	Supv	-	8.4/194	19.0/227	19.1/232
	-	-	8.1/150	18.9/201	18.8/201
	1		8.1/139	18.9/188	18.7/188
	3g		8.0/133	18.8/176	18.7/184
3g	Supv	-	8.1/179	19.0/213	18.7/215
	3g	**A**	8.1/128	18.9/176	18.7/179
		B	8.0/120	18.9/168	18.7/171
		C	8.0/126	19.0/180	18.7/180
		D	7.9/118	18.8/166	18.5/169

Table 4: CER and PP performance of ML adapted 4-gram LMs on bn06, bc05 and eval06.

Performance of combining training set nPP and test adapted weights using the methods proposed in section 3 are shown in the last four lines of table 4. The output from the nPP system in table 2 was used as the adaptation supervision, as is shown in the 5th line of table 4. The two-stage MAP estimation (option **A**) and log-linear composition approaches (option **C**) gave similar performance. Using the a two-stage MAP estimation with union (option **B**), further PP reduction was obtained but the CER gains were minimum. The weighted log-linear composition and union approach (option **D**) gave the best performance among the four. More than 30 points of PP reduction (17%-22% rel) and 0.1%-0.3% absolute CER reduction were obtained over the adapted baseline using context free weights (2nd line of table 4). These results suggest sufficient coverage of contexts in test set supervision is important when adapting LMs using context dependent interpolation.

ML 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 is now interesting to investigate the performance of MBR discriminative adaptation. These are shown in table 5. The top 1000 hypotheses were extracted from the lattices generated by the unadapted baseline system (also in first line of table 5) for MBR self-adaptation and the smoothing constant D set in the same way as described in [4]. Performance of three standard MBR adapted systems with context free or dependent weights are shown from the 2nd to 4th line. The 3-gram weight MBR system gave the best adaptation performance. Statistically significant CER reductions of 0.4% on bn06, bcdev05 and 0.5% on eval06 over the unadapted baseline were obtained. Using a weighted log-linear composition and union based approach (option **D**) to combine with the nPP system of table 2 gave further CER improvement of 0.2% on bc05. The CER gains over the adapted baseline using global weights are 0.2% on bn06 and 0.3% on bc05 and eval06. The total CER gains over the unadapted baseline system are 0.5% (6% rel) on bn06, 0.4% on bc05 and 0.6% on eval06, all being statistically significant.

Context Prior	Context Adapt	Wgt Com	CER% bn06	bc05	eval06
-	Supv	-	8.4	19.0	19.1
	-	-	8.1	18.9	18.8
	1g		8.1	18.7	18.6
	3g		8.0	18.6	18.6
3g	3g	**D**	7.9	18.6	18.5

Table 5: CER performance of MBR adapted 4-gram LMs on bn06, bc05 and eval06.

5. Conclusion

Context dependent LM adaptation by combining training and test set information under a discriminative framework was investigated in this paper. Experimental results on a state-of-the-art large vocabulary speech recognition task suggest that the proposed method may be useful for speech recognition. Future research will focus on using discriminative training techniques in both model interpolation and adaptation stages. Continuous forms of history weighting function will also be investigated.

6. References

[1] I. Bulyko, M. Ostendorf & A. Stolcke. "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures", in *Proc. HLT'03.*

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

[3] 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.

[4] 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.

[5] X. Liu, M. J. F. Gales & P. C. Woodland (2008). Context Dependent Language Model Adaptation, in *Proc. Interspeech'08*, Brisbane.

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