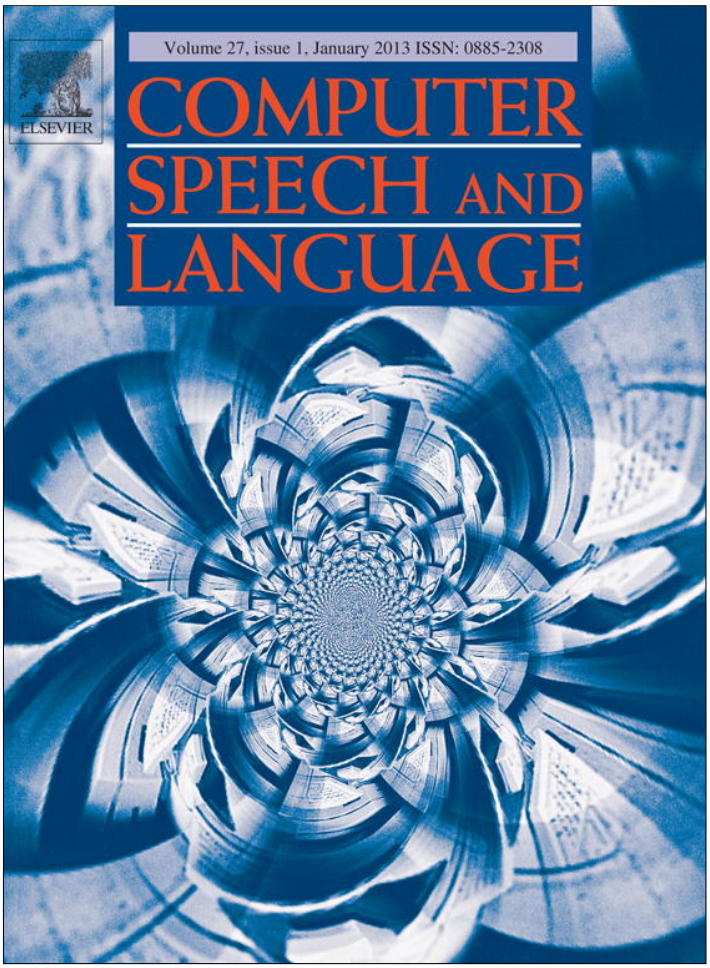

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 301–321

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Use of contexts in language model interpolation and adaptation☆

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

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England, United Kingdom

Received 22 July 2011; received in revised form 11 June 2012; accepted 12 June 2012
Available online 27 June 2012

Abstract

Language models (LMs) are often constructed by building multiple individual component models that are combined using context independent interpolation weights. By tuning these weights, using either perplexity or discriminative approaches, it is possible to adapt LMs to a particular task. This paper investigates the use of context dependent weighting in both interpolation and test-time adaptation of language models. Depending on the previous word contexts, a discrete *history weighting function* is used to adjust the contribution from each component model. As this dramatically increases the number of parameters to estimate, robust weight estimation schemes are required. Several approaches are described in this paper. The first approach is based on MAP estimation where interpolation weights of lower order contexts are used as smoothing priors. The second approach uses training data to ensure robust estimation of LM interpolation weights. This can also serve as a smoothing prior for MAP adaptation. A *normalized perplexity* metric is proposed to handle the bias of the standard perplexity criterion to corpus size. A range of schemes to combine weight information obtained from training data and test data hypotheses are also proposed to improve robustness during context dependent LM adaptation. In addition, a minimum Bayes' risk (MBR) based discriminative training scheme is also proposed. An efficient weighted finite state transducer (WFST) decoding algorithm for context dependent interpolation is also presented. The proposed technique was evaluated using a state-of-the-art Mandarin Chinese broadcast speech transcription task. Character error rate (CER) reductions up to 7.3% relative were obtained as well as consistent perplexity improvements.

1. Introduction

A crucial component in automatic speech recognition (ASR) systems is the language model. Back-off *n*-gram models remain the dominant language modelling approach for state-of-art ASR systems (Katz, 1987). In these systems language models are often constructed by combining component *n*-gram models trained on a diverse collection of text sources prior to probability interpolation. Individual data sources will be more appropriate depending on the task, for example, broadcast news or conversational telephone speech. To reduce the mismatch between the interpolated model and the target domain of interest, interpolation weights may be tuned by minimizing the perplexity on some held-out data similar to the target domain (Jelinek and Mercer, 1980; Kneser and Steinbiss, 1993; Iyer et al., 1994; Bahl et al., 1995; Rosenfeld, 1996, 2000; Jelinek, 1997; Clarkson and Robinson, 1997; Kneser and Peters, 1997; Seymore and Rosenfeld, 1997; Iyer and Ostendorf, 1999). These weights indicate the "usefulness" of each source for a particular task. To further improve robustness to varying styles or tasks, unsupervised test-set adaptation, for example, to a particular

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author. Tel.: +44 1223 766512; fax: +44 1223 332662.
E-mail addresses: xl207@cam.ac.uk (X. Liu), mjfg@eng.cam.ac.uk (M.J.F. Gales), pcw@eng.cam.ac.uk (P.C. Woodland).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.06.004

broadcast show, may be used (Della Pietra et al., 1992; Bulyko et al., 2012, 2007; Federico, 1999, 2003; Gildea and Hofmann, 1999; Chen et al., 2001; Mrva and Woodland, 2004, 2006; Chien et al., 2005; Tam and Schultz, 2005; Liu et al., 2007, 2008, 2009, 2010). As directly adapting *n*-gram word probabilities is impractical on limited amounts of data, standard adaptation schemes only involve updating one single, context independent interpolation weight for the component models (Iyer et al., 1994; Rosenfeld, 1996; Clarkson and Robinson, 1997; Seymore and Rosenfeld, 1997; Iyer and Ostendorf, 1999; Mrva and Woodland, 2006).

There are two major issues with this standard adaptation scheme. First, the diversity among data sources manifests itself in a wide range of factors, including source of collection, epoch, genre, modelling resolution and robustness, topic coverage and style. The precise nature of each source is jointly determined by a combination of these factors. Some of these factors are adequately modelled on a higher level using global, context *independent* weights, based on, for example, source of collection, epoch or genre. However, others factors, such as *n*-gram modelling resolution and generalization, topic coverage and style, can also affect the contribution of sources on a local, *context dependent* basis. Thus the usefulness of a particular source for a domain may vary depending on the word context for both model interpolation and adaptation. Using global weights takes no account of this local variability. Hence, it is preferable to increase the modelling resolution of weight parameters by adding context information (Bulyko et al., 2012; Hsu, 2007; Liu et al., 2008). In previous research, context dependent interpolation weights were tuned on a development set (Bulyko et al., 2012). Little attempt was made to further adapt the language model to a finer domain, for example, a particular broadcast show or conversation. Due to data sparsity, only a small number of parameters can be estimated. Hence, the performance improvement from additional context information in model interpolation is often limited. For example, in Bulyko et al. (2012) LM interpolation weights were shared among class-based *n*-gram histories derived from part-of-speech (PoS) information. Small error rate improvements of 0.1–0.4% absolute were reported on a conversational telephone speech task where the baseline system gave an error rate above 38%. In order to robustly estimate a large number of interpolation weights that generalize well to a wide range of tasks, it would be preferable to use empirically available training data.

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 training techniques (Och and Ney, 2012; Bulyko et al., 2007; Liu et al., 2007) to estimate interpolation weight parameters. These schemes do not rely on incorrect modelling assumptions and explicitly aim at reducing the underlying error cost function. In particular the minimum Bayes' risk (MBR) criterion provides a flexible framework that can generalize to a wide range of error cost functions (Kaiser et al., 2012; Povey and Woodland, 2002; Doumpiotis and Byrne, 2005).

To address the first issue, this paper investigates the use of context dependent interpolation in both training and test-time unsupervised adaptation of language models. Depending on the previous word contexts, a discrete *history weighting function* is used to dynamically adjust the contribution from each component model. As this dramatically increases the number of parameters to estimate, robust weight estimation schemes are required. Several approaches are described in this paper. The first is based on *maximum a posteriori* (MAP) estimation where interpolation weights of lower order contexts are used as smoothing priors. The second approach uses training data to ensure robust estimation for a general form of context dependent LM interpolation. This can also serve as a MAP smoothing prior with higher context resolution. An important issue with this method is to handle the bias to corpus size. An inverse corpus size weighted form of perplexity, *normalized perplexity*, is proposed to address this issue. The third approach combines weights estimated from the training data and test data hypotheses to improve robustness in context dependent LM adaptation. For unseen contexts a weight back-off scheme is used.

In order to reduce the error cost function mismatch between model estimation and performance evaluation, MBR based discriminative training schemes are also proposed to improve context dependent LM adaptation. In order to flexibly support a wide range of LM interpolation and adaptation configurations, an efficient weighted finite state transducer (WFST) based on-the-fly decoding algorithm is also presented. Performance is evaluated on a state-of-the-art large vocabulary Mandarin Chinese broadcast transcription task. Consistent perplexity and error rate improvements were obtained over baseline systems using global, context independent weights.

2. Language model interpolation and adaptation

A common approach for LM adaptation is to adjust the context independent, linear interpolation weights for a mixture model (Rosenfeld, 1996; Clarkson and Robinson, 1997; Seymore and Rosenfeld, 1997; Iyer and Ostendorf,

1999; Mrva and Woodland, 2006). For word based n-gram models, the log probability of the L word sequence $\mathcal{W} = \langle w_1, w_2, \ldots, w_i, \ldots, w_L \rangle$, is given by

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

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, $\langle w_{i-1}, w_{i-2}, \ldots, w_{i-n+1} \rangle$. For language modeling in current ASR systems, two basic forms of probability interpolation are available: a linear or log-linear interpolation of component models (Jelinek and Mercer, 1980; Rosenfeld, 1996, 2000; Jelinek, 1997; Darroch and Ratcliff, 1972; Rosenfeld et al., 2001; Och and Ney, 2012). These in turn are instances of mixtures of experts (MoE) and products of experts (PoE) in machine learning literature (Hinton, 2002).

2.1. Linear and log-linear interpolation

The linearly interpolated word probability is computed as,

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

where λ_m is the global weight for the mth component model. As a *union* of all the individual experts, linear interpolation tends to give a broader distribution than individual components alone. Hence, this form of model combination may help overcome the sparsity issue when training individual component models and thus improve generalization.

In contrast, a log-linear interpolation provides an *intersection* of individual experts. It yields a high likelihood only when all component models agree. Using the same example as above, this is given by,

$$P(w_i|h_i^{n-1}) = \frac{1}{Z_{h_i^{n-1}}} \exp\left(\sum_m \lambda_m \ln P_m(w_i|h_i^{n-1})\right) \tag{3}$$

where $Z_{h_i^{n-1}}$ is a normalization term to ensure that the interpolated probability to be a valid probability distribution. However, the exact computation of the normalization term for general forms of log-linear models is non-trivial, and analytical solutions are available only for certain forms of probability functions. The normalization term may be ignored when interpolation is performed at the complete word sequence level under a discriminative framework such as *maximum entropy* models and *logistic regression* (Darroch and Ratcliff, 1972; Rosenfeld et al., 2001; Och and Ney, 2012).

2.2. Multi-level LM interpolation

The precise nature of the component language models determines which of the above two combination schemes is more appropriate during model interpolation. For example, when building word level LMs, in order to improve context coverage and generalization, a linear interpolation between component LMs trained over a diverse set of text sources can be used. It was also found in previous research that when introducing additional sub-word level linguistic constraints to increase discrimination, word and syllable level LMs can be combined in a log-linear fashion (Hieronymus et al., 2009; Liu et al., 2010). In order to obtain a good balance between generalization and discrimination, it is also possible to leverage from both linear and log-linear forms of model combination. A multi-level LM interpolation represented by a product between a mixture of experts, or equivalently a log-linear combination of linearly interpolated language models, may be considered.

2.3. Maximum likelihood estimation of interpolation weights

Assuming a "sufficient" amount of training data is available, and a strong correlation between perplexity and error rate exists, a perplexity based, or equivalently maximum likelihood (ML), estimation scheme is often used for optimizing global interpolation weights. The objective is given by,

$$\mathcal{F}_{\mathsf{PP}}(\boldsymbol{\lambda}) = \exp\left\{-\frac{\ln P(\mathcal{W})}{L}\right\} = \exp\left\{-\frac{1}{L}\sum_{i=1}^{L}\ln\left(\sum_{m=1}^{M}\lambda_m P_m(w_i|h_i^{n-1})\right)\right\} \tag{4}$$

which is equivalent to maximizing the log-likelihood of the entire word sequence, $\ln P(\mathcal{W})$. Here $\mathcal{W}$ is the held-out data for interpolation tuning, or supervision in case of model adaptation. The optimal linear interpolation weight for the mth component model can be derived by,

$$\hat{\boldsymbol{\lambda}}^{\mathsf{ML}} = \underset{\boldsymbol{\lambda}}{arg\max}\left\{\sum_{i=1}^{L}\ln\left(\sum_{m=1}^{M}\lambda_m P_m(w_i|h_i^{n-1})\right)\right\} \tag{5}$$

Under a constraint such that $0<\lambda_m<1$ and $\sum_m\lambda_m = 1$, the weight parameters may be iteratively re-estimated as,

$$\hat{\lambda}_m^{\mathsf{ML}} = \frac{\mathcal{C}_m^{\mathsf{ML}}(\mathsf{null})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(\mathsf{null})} \tag{6}$$

where the ML context *independent* statistics, $\mathcal{C}_m^{\mathsf{ML}}(\mathsf{null})$ are defined as,

$$\mathcal{C}_m^{\mathsf{ML}}(\mathsf{null}) = \tilde{\lambda}_m \left.\frac{\partial \ln P(\mathcal{W})}{\partial \lambda_m}\right|_{\boldsymbol{\lambda}=\tilde{\boldsymbol{\lambda}}} \tag{7}$$

$\tilde{\lambda}$ represents the current weight estimates, and the derivative term is computed as,

$$\frac{\partial \ln P(\mathcal{W})}{\partial \lambda_m} = \sum_{i=1}^{L}\frac{P_m(w_i|h_i^{n-1})}{\sum_m \lambda_m P_m(w_i|h_i^{n-1})} \tag{8}$$

Note that the above notation for sufficient weight statistics $\mathcal{C}_m^{\mathsf{ML}}(\cdot)$ is not normally used for standard language model interpolation with context independent weights. In this paper it provides additional flexibility to restrict the statistics to be accumulated for, for example, a particular word history context. This form of notation for interpolation weight statistics will be extensively used in the rest of the paper.

3. Discriminative language model interpolation and adaptation

As discussed in Section 1, for current speech recognition systems the correlation between perplexity and error rate is fairly weak. Hence, it would be preferable to employ discriminative training techniques that explicitly aim at reducing the underlying error rate cost function, such as the MBR criterion (Kaiser et al., 2012; Povey and Woodland, 2002; Doumpiotis and Byrne, 2005) to estimate context dependent interpolation weights. The MBR criterion is expressed as the expected recognition error of an ASR system on a sequence of speech observations, $\mathcal{O}$. It is computed by summing over the cost function contribution from all possible hypotheses $\{\mathcal{W}\}$, weighted by their posterior probabilities, $P(\mathcal{W}|\mathcal{O})$. Taking the global, context independent interpolation weight parameters as an example, they are optimized by,

$$\begin{aligned}\hat{\boldsymbol{\lambda}}^{\mathsf{MBR}} &= \underset{\boldsymbol{\lambda}}{arg\min}\{\mathcal{F}_{\mathsf{MBR}}(\mathcal{O};\boldsymbol{\lambda})\} = \underset{\boldsymbol{\lambda}}{arg\min}\left\{\sum_{\mathcal{W}}P(\mathcal{W}|\mathcal{O})\mathcal{L}(\mathcal{W},\mathcal{W}_{\mathsf{ref}})\right\}\\ &= \underset{\boldsymbol{\lambda}}{arg\min}\left\{\sum_{\mathcal{W}}\frac{P(\mathcal{O}|\mathcal{W})\prod_i\left(\sum_m\lambda_m P_m(w_i|h_i^{n-1})\right)}{\sum_{\mathcal{W}}P(\mathcal{O}|\mathcal{W})\prod_i\left(\sum_m\lambda_m P_m(w_i|h_i^{n-1})\right)}\mathcal{L}(\mathcal{W},\mathcal{W}_{\mathsf{ref}})\right\}\end{aligned} \tag{9}$$

where $\mathcal{L}(\mathcal{W},\mathcal{W}_{\mathsf{ref}})$ denotes the defined recognition error rate measure of hypothesis $\mathcal{W}$ against the reference hypothesis $\mathcal{W}_{\mathsf{ref}}$. A variety of forms of cost function, such as word or sub-word level error rates, may be used depending on the

underlying evaluation metric being considered. This provides more flexibility, compared with other discriminative criteria, such as maximum mutual information (MMI) (Normandin, 1991), as the loss function can be task dependent and not necessarily restricted to one particular form. By definition if $\mathcal{W}_{\text{ref}}$ is the correct transcription MBR adaptation will be performed in supervised mode.

The Extended Baum-Welch (EBW) algorithm may be used to optimize the MBR criterion. It provides an efficient iterative optimization scheme for a family of rational objective functions, including MBR (Gopalakrishnan et al., 1991). For global linear interpolation weights under a sum-to-one constraint, the re-estimation formula is given by,

$$\hat{\lambda}_m^{\text{MBR}} = \frac{\mathcal{C}_m^{\text{MBR}}(\text{null})}{\sum_m \mathcal{C}_m^{\text{MBR}}(\text{null})} \tag{10}$$

where the discriminative context *independent* statistics, $\mathcal{C}_m^{\text{MBR}}(\text{null})$, are computed as

$$\mathcal{C}_m^{\text{MBR}}(\text{null}) = \tilde{\lambda}_m \left.\frac{\partial \mathcal{F}_{\text{MBR}}(\mathcal{O};\lambda)}{\partial \lambda_m}\right|_{\lambda=\tilde{\lambda}} + D \tag{11}$$

$\tilde{\lambda}$ is the current weight estimate, and D a tunable regularization constant controlling the convergence speed. Following the MBR criterion defined as above, the partial derivative in the above may be re-expressed as Liu et al. (2007),

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

where the first term can be derived as the following,

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

The second term in Eq. (12) is independent of the acoustic model distribution $p(\mathcal{O}|\mathcal{W})$, hence one may write

$$\frac{\partial \ln p(\mathcal{O},\mathcal{W})}{\partial \lambda_m} = \frac{\partial \ln P(\mathcal{W})}{\partial \lambda_m} = \sum_{i=1}^{L} \frac{P_m(w_i|h_i^{n-1})}{\sum_m \lambda_m P_m(w_i|h_i^{n-1})} \tag{14}$$

which is effectively identical to the sufficient statistics required by the perplexity based optimization given in Eq. (8).

When using the MBR criterion to train context dependent interpolation weights, the MAP estimation schemes given in Eqs. (17) and (18), together with the weight set combination schemes proposed in Section 5.3 can also be used for comparable MBR statistics, $\mathcal{C}_m^{\text{MBR}}(\cdot)$. One exception is the discriminative estimation of interpolation weights on the training data. This would require all data sources to have confusable word sequences explicitly generated. It is a non-trivial problem for training sources other than audio transcriptions. In this paper, MBR estimation is only used for test-time LM self-adaptation. The use of training data for general model interpolation is only considered using normalized perplexity or ML.

4. Context dependent interpolation and adaptation

As discussed above, when global weights are assigned to component n-gram language models, no account is taken of the surrounding contexts. In order to incorporate context information, a more general form is to use a context dependent *history weighting function*, $\boldsymbol{\phi}(h)$. Using an n-gram context history, the interpolated word probability in Eq. (2) becomes (Liu et al., 2008, 2009),

$$P(w_i|h_i^{n-1}) = \sum_m \boldsymbol{\phi}_m(h_i^{n-1}) P_m(w_i|h_i^{n-1}) \tag{15}$$

where $\boldsymbol{\phi}_m(h_i^{n-1})$ is the mth component weight vector for n-gram history h_i^{n-1}. The Markov chain assumption used in the component n-gram models is made such that the interpolation weights for word w_i only depend on the preceding $n-1$ words. These context dependent weights are also constrained to be positive and sum-to-one. A history weighting function could have either a discrete or a continuous form. In this paper only discrete forms of history weighting function are considered.

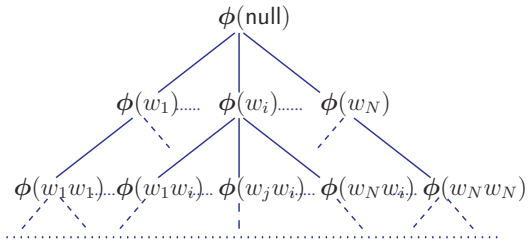

Fig. 1. Hierarchy of context dependent interpolation weights for back-off tri-gram language models with a history of two words maximum.

Discrete history weighting functions are effectively look-up tables of n-gram context dependent weights. In contrast to the traditional context independent, global weighting of component models given in Eq. (2), each distinct history can have its own interpolation weight vector. Such tying of interpolation weights at history context level may be viewed as an alternative to directly re-estimating the n-gram probabilities for LM adaptation. Compared to other LM adaptation methods such as latent semantic analysis (LSA) (Bellegarda, 2000; Brants, 2005) and latent dirichlet allocation (LDA) (Blei et al., 2003), which do not explicitly consider the exact word order in the history context, or minimum discrimination information (MDI) (Federico, 1999, 2003), which only adjusts unigram probabilities to match the adaptation data, the above form of context dependent weighting can more precisely model the change in word history during LM adaptation. However, the number of parameters to estimate will dramatically increase as the length of history context grows. For contexts observed in the training or adaptation data, a tree structured hierarchy of history dependent interpolation weights can be used. An example is shown in Fig. 1 for a back-off tri-gram LM. For history contexts that are not seen in the training or adaptation data, this tree structure of contexts allows an efficient back-off strategy to be used in the same way as standard n-gram language models. This hierarchy will be extensively used in the rest of this paper.

$$\boldsymbol{\phi}^{\mathsf{bo}}(h_i^{n-1}) = \begin{cases} \boldsymbol{\phi}(h_i^{n-1}) & \text{if } \exists\, \boldsymbol{\phi}(h_i^{n-1}) \\ \boldsymbol{\phi}(h_i^{n-2}) & \text{else if } \exists\, \boldsymbol{\phi}(h_i^{n-2}) \\ \cdots & \cdots \\ \boldsymbol{\phi}(\mathsf{null}) & \text{otherwise} \end{cases} \tag{16}$$

where $\boldsymbol{\phi}^{\mathsf{bo}}(h_i^{n-1})$ represents the back-off estimate of the mth component weight vector for n-gram history h_i^{n-1}. The above back-off recursion will simplify to the global, context independent interpolation weight, $\boldsymbol{\phi}(\mathsf{null})$, if the preceding word, w_{i-1}, is not observed in the data. These discrete weight parameters can be ML trained using the re-estimation algorithm discussed in Section 2. The difference between the two approaches is that the context dependent form of statistics $\{\mathcal{C}_m(h_i^{n-1})\}$ are now required in the update formulae of Eq. (6). Because the re-estimated weight parameters are constrained to be positive and sum-to-one, no normalization term is required for the back-off process. This is simpler than standard word or class-based n-gram models, in which case context dependent back-off weights need to be explicitly computed to satisfy the sum-to-one constraint. Since the number of weight parameters increases exponentially with context length, robust weight estimation schemes are required when only a limited amount of training data is available.

5. Robust estimation of context dependent weights

As discussed in Section 4, when the history length grows, the number of context dependent interpolation weights that need to be estimated increases exponentially. During model tuning or test time adaptation, often only a limited amount of data is available. This data sparsity issue is particularly important and must be addressed. In this section, several types of techniques are proposed to handle this issue. The first approach is based on MAP estimation where interpolation weights of lower order contexts are used as smoothing priors. The second approach uses training data to ensure robust estimation for context dependent LM interpolation. These interpolation weights can also be used as MAP smoothing priors for test-time LM adaptation. The third approach uses an appropriate combination between the

weights estimated from the training data and test data hypotheses to improve robustness in context dependent LM adaptation. Both MAP and log-linear composition based combination methods are presented.

5.1. MAP estimation

One common approach to address the robustness issue is to use maximum *a-posteriori* (MAP) estimation. This is a general approach and can be applied to improve robustness for both the ML and discriminative weight estimation schemes discussed in Sections 2.3 and 3. Taking the perplexity or ML based adaptation is as an example, this is given by

$$\hat{\boldsymbol{\phi}}_m^{\mathsf{MAP}}(h_i^{n-1}) = \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau \boldsymbol{\phi}_m^{\mathsf{Pr}}(h_i^{n-1})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau} \tag{17}$$

where $\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1})$ is context dependent ML statistics for history context h_i^{n-1} given in Eq. (7), and τ controls the contribution from weight prior, $\boldsymbol{\phi}_m^{\mathsf{Pr}}(h_i^{n-1})$.

One key issue with MAP estimation is the choice of smoothing prior. The global, context independent weights, $\boldsymbol{\phi}(\mathsf{null})$, may be used (Liu et al., 2008). In order to introduce more context information, rather than completely backing off to the context independent weights, hierarchical smoothing using weights of lower order contexts may also be considered, as inspired by interpolated Kneser–Ney smoothing of n-gram models (Chen and Goodman, 1999). Hierarchical smoothing priors can be used to improve robustness for both the above mentioned ML and discriminative weight estimation schemes discussed in Sections 2.3 and 3. Taking the perplexity based estimation as an example, this is given by

$$\hat{\boldsymbol{\phi}}_m^{\mathsf{hier}}(h_i^{n-1}) = \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau \hat{\boldsymbol{\phi}}_m^{\mathsf{hier}}(h_i^{n-2})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau} \tag{18}$$

When adapting LMs using context dependent interpolation, the above form of hierarchical smoothing provides a dynamic, in-domain prior estimated on the test data hypotheses. However, such a prior may be less informative due to reduced context resolution. It can also be sensitive to errors in the supervision hypotheses. Alternatively, a different type of prior with finer context dependent modelling may be obtained from the training data.

5.2. Model interpolation using training data

Large amounts of text data, for example, billions of words, are often used to train component n-gram models for state-of-the-art large vocabulary speech recognition systems (Liu et al., 2007). These can be used to ensure robust weight estimation for a more informative weight prior with a longer context span. Because the prior is estimated over a wide range of data types found in the training data, this weight prior itself may also be used for building a domain neutral interpolated language model.

When using training data from multiple text sources, for both context independent and dependent LM interpolation given by Eqs. (2) and (15), the sufficient statistics in Eq. (8) for re-estimation will be accumulated over every word within each sentence of each individual text source. The statistics will be dominated by large sized, potentially non-useful corpora, and thus introduce a bias. This is a fundamental issue that can affect the performance of the interpolated model and its use as a smoothing prior for MAP adaptation. Hence, such a bias to the corpus size must be addressed. The approach proposed in this paper is to use a corpus length normalization scheme, which ensures that the word count contribution from each corpus is the same. The training data log-probability given in Eq. (15) is modified as,

$$\ln P_{\mathsf{norm}}(\mathcal{W}) = \sum_m \frac{L}{L_m} \sum_{i=1}^{L_m} \ln P(w_i | h_i^{n-1}) \tag{19}$$

where L_m is the total number of words in the mth corpus, and the total number of words in all corpora is defined as $L = \sum_m L_m$. The normalized perplexity (nPP) measure is computed using the above as

$$\mathcal{F}_{\mathsf{nPP}} = \exp\left\{-\frac{\ln P_{\mathsf{norm}}(\mathcal{W})}{\sum_m L}\right\} \tag{20}$$

Using the nPP criterion, weights are determined by the average word log-probability from each data corpus. As the dependency upon word counts is removed, the bias to larger sized corpora can thus be handled.

As discussed in Section 1, the variability among data sources and their contribution are jointly determined by a combination of multiple attributes. Some factors, for example, epoch and genre, may be sufficiently modelled using global, context independent weights. These context independent weights measure the contribution of data sources on a higher level. Other factors such as modelling resolution, topic coverage and style, require local, context dependent weighting. These weights indicate that for each context some sources are more appropriate than others. Using the nPP criterion, the interpolation weights at both levels can be estimated.

In practice the global level contribution among sources is often further increased by taking conscious decisions when building component models. If certain sources such as the acoustic transcriptions, are known to be useful for the domain of interest, a bias to these components of the same genre may be introduced during LM construction. When lower cut-offs[1] are used for these sources, the associated component models will have high probabilities on their training data compared to others built with higher cut-off settings. Similarly if robust discounting schemes are used then these models will also generalize well on other data.

In order to improve robustness for rare contexts, the hierarchical smoothing based MAP estimation given in Eq. (18) can also be used for nPP statistics. This is given by,

$$\hat{\boldsymbol{\phi}}_m^{\mathsf{nPP}}(h_i^{n-1}) = \frac{C_m^{\mathsf{nPP}}(h_i^{n-1}) + \tau\hat{\boldsymbol{\phi}}_m^{\mathsf{nPP}}(h_i^{n-2})}{\sum_m C_m^{\mathsf{nPP}}(h_i^{n-1}) + \tau} \tag{21}$$

As the statistics are obtained from a wide range of tasks found in the training data, the estimated weights can be used to build a domain neutral interpolated LM. Furthermore, they can be also employed as smoothing priors for test-time adaptation of context dependent interpolation weights using Eq. (17).

5.3. Weight set combination

As discussed previously in Sections 5.1 and 5.2, 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. They provide either domain neutral, longer context based weights, or in-domain, shorter context based ones:

- *Training*: data nPP weights estimated using the hierarchical smoothing given in Eq. (21). They provide richer context information and finer modelling resolution, but potentially a larger mismatch with the target domain during LM adaptation.
- *Test*: data adapted weights using Eq. (18) with 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 Eq. (16) is necessary. This will result in reduced modelling resolution.

Given the different nature of these two sources of information, it is preferable to appropriately combine them for context dependent LM adaptation. 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 technique: MAP estimation and log-linear weight combination. Within each category, it is also possible to further supplement the adapted weights of contexts obtained from the training data with those of contexts newly observed in the test data hypotheses used for adaptation supervision. These additional contexts may carry additional information from the target domain for

[1] Cut-offs determine the count threshold for an n-gram to be retained in a model. Hence a small cut-off value will retain a large number of n-gram entries.

adaptation. Hence, it is also interesting to estimate distinct interpolation weights for them, instead of relying on weight back-off through the training data context hierarchy using Eq. (16). Note that the hierarchical smoothing of Eq. (18) effectively is based on a lower order context based weight prior. This was also used for robust nPP estimation of Eq. (21). For clarity in the rest of this paper the term "prior" is reserved and exclusively refers to nPP weights estimated from the training data. The combination schemes proposed in this section can be applied to improve robustness for both the ML and discriminative weight estimation schemes discussed in Sections 2.3 and 3. For simplicity, perplexity or ML based adaptation is taken as an example here.

5.3.1. MAP estimation

Initially nPP based LM interpolation is performed. Contexts are extracted from the training data. To improve robustness for rare contexts, their weights are MAP estimated using hierarchical smoothing as in Eq. (21). During test-time LM adaptation, these nPP estimated context dependent weights are used as a smoothing prior. The final adapted mth component weight of history context h_i^{n-1} is thus given by

$$\hat{\boldsymbol{\phi}}_m^{\mathsf{comb}}(h_i^{n-1}) = \hat{\boldsymbol{\phi}}_m^{\mathsf{MAP}}(h_i^{n-1}) = \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau\hat{\boldsymbol{\phi}}_m^{\mathsf{nPP}}(h_i^{n-1})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau} \tag{22}$$

5.3.2. MAP estimation and union

In option A all contexts are exclusively obtained from the training data. As previously discussed, new contexts uniquely observed in the test set adaptation supervision may carry additional useful information from the target domain of interest. Hence, it is interesting to also estimate their associated weights. There is a choice of the smoothing prior for these newly acquired contexts from the supervision hypothesis. Two forms of weight prior may be used. The first one is a back-off estimate of the static nPP prior obtained using Eq. (16). The second uses a hierarchical dynamic smoothing prior based on "parental" contexts as in Eq. (18). Both have a reduced modelling resolution due to the back-off to a lower order context. Between the two, the dynamic prior contains more direct information of the target domain for LM adaptation. Hence, it is used for new contexts found in the supervision. Thus the MAP re-estimation formulae given in Eq. (22) is extended to

$$\hat{\boldsymbol{\phi}}_m^{\mathsf{comb}}(h_i^{n-1}) = \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau\hat{\boldsymbol{\phi}}_m^{\mathsf{nPP}}(h_i^{n-1}) + \tau_h\hat{\boldsymbol{\phi}}_m^{\mathsf{hier}}(h_i^{n-2})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau + \tau_h} \tag{23}$$

where τ_h controls the contribution from dynamic hierarchical weight prior. For contexts obtained from the training data, τ_h is set as uninformative and only the nPP prior is used, thus $\tau > 0$ and $\tau_h = 0$. For new contexts found in the supervision hypotheses, the nPP prior becomes uninformative while only the dynamic hierarchical prior is used, thus $\tau = 0$ and $\tau_h > 0$. The aim of alternating between the two priors is to achieve a good balance between the nPP prior's modelling resolution for observed longer span contexts and the dynamic prior's domain relevance. This approach can be viewed as an extension to option A. The context tree of Fig. 1 is effectively expanded by adding more nodes that represent the newly observed histories in the supervision.

5.3.3. Log-linear composition

MAP estimation may be viewed as a weighted linear interpolation between, for example, an ML estimate and its smoothing prior. For context dependent LM adaptation, there are two issues with this approach. First, training data extracted contexts that are unavailable in the supervision will back-off to a domain neutral nPP prior containing minimum information about the test data. Second, due to the nature of linear interpolation, test set weights that are MAP adapted to incorrect supervision can retain certain mis-ranking of component LMs. To address these issues, an alternative is to use a log-linear combination between the training and test data estimated weight sets. After a normalization to satisfy the sum-to-one constraint, the final combined weights are,

$$\hat{\boldsymbol{\phi}}_m^{\mathsf{comb}}(h_i^{n-1}) = \frac{\exp\left\{\alpha \ln \hat{\boldsymbol{\phi}}_m^{\mathsf{nPP}}(h_i^{n-1}) + \ln \hat{\boldsymbol{\phi}}_m^{\mathsf{hier}}(h_i^{n-1})\right\}}{\sum_m \exp\left\{\alpha \ln \hat{\boldsymbol{\phi}}_m^{\mathsf{nPP}}(h_i^{n-1}) + \ln \hat{\boldsymbol{\phi}}_m^{\mathsf{hier}}(h_i^{n-1})\right\}} \tag{24}$$

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

Using a log-linear combination, all contexts extracted from the training data to estimate the nPP prior will be adapted. Rather than being left unadapted in MAP estimation as in Eq. (22), weights of unseen contexts will be also adapted using a back-off estimate obtained from the test data using Eq. (16). Furthermore, a log-linear combination can also reject the component LM weighting obtained from erroneous supervision hypotheses that are very different from the training set nPP prior. Therefore it can improve the robustness of weight combination.

5.3.4. *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 in the test data supervision and therefore completely backs off to the context independent, global weights, Eq. (24) ($\alpha = 1.0$) 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 Eq. (24). In common with B, weights of contexts uniquely observed in the test set supervision are also added. Again the goal is to leverage both the nPP prior's high modelling resolution for longer span contexts and the dynamic prior's domain relevance for LM adaptation. 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. Note that it is also possible to perform a "full" composition for all contexts observed both in training and testing by fixing $\alpha = 1.0$. However, in practice this setting was found to lead to a slight performance degradation.

6. Interpolated LMs and weighted finite state transducers

As discussed in Section 1, in current ASR systems language models are often constructed by training *n*-gram components models on data from a set of diverse sources. Interpolated LMs with context independent interpolation weights are normally built using special purpose tools, for example, the SRILM or HTK toolkit (Stolcke, 2002; Young et al., 2009). In order to capture local variations in modelling resolution, generalization, topic coverage and style among component LMs, the history context dependent form of LM interpolation and adaptation introduced in Section 4 can be used. To incorporate more linguistic constraints, it is also possible to train and combine LMs that model different types of unit sequences, for example, syllables and words (Hieronymus et al., 2009). These techniques often require extensive software changes.

An alternative, and more general, approach proposed here is to interpolate component language models using weighted finite state transducers (WFSTs) (Mohri, 1997; Mohri and Riley, 1998; Mohri et al., 1998, 2000; Liu et al., 2010). As this approach is entirely based on well-defined WFST operations, only a minimal change to decoding tools is required. It is highly flexible and can be used for a wide range of LM combination configurations. It not only supports the use of global, context independent weights in LM combination, but also the more general case when context dependent weights are employed. Thus context dependent LM interpolation and adaptation can be conveniently implemented. Unless otherwise stated, *tropical semi-ring* based WFSTs (Mohri, 1997; Mohri and Riley, 1998; Mohri et al., 1998, 2000) are considered in this paper.

A WFST is a finite state machine that carries weights such as log-probabilities that accumulate linearly along paths within a directed graph to each pair of input and output symbol sequences. A set of classic finite automata operations to combine, optimize and compact WFSTs during search are available. Many types of modelling information used in speech recognition systems, such as the HMM topology, lexicon, *n*-gram models and the context dependent weight models considered here, involve a stochastic finite-state mapping between symbol sequences. WFSTs provide a generic and well-defined framework to represent them. More precisely, *n*-gram language models and context dependent interpolation weights can be represented by weighted finite state acceptors (WFSA). These are special cases of WFSTs when the input and output symbol sequences are identical. As discussed in Section 2.1, component LMs can be combined using both linear and log-linear forms of model interpolation, or a combination of the two as in multi-level LM interpolation. Using a WFST based representation, these combination schemes may be efficiently implemented using the *union*, *composition* or *intersection* operations, or a combination of them.

6.1. Linear model combination

For both context independent and dependent linear language model combination, the WFST representation of the final interpolated LM may be derived using a component level *composition* between the *n*-gram and interpolation weight transducers prior to a final *log semi-ring* based *union* operation performed locally at the *n*-gram level. This is given by,

$$\mathcal{L} = (\mathcal{L}_G^{(1)} \circ \mathcal{L}_{\boldsymbol{\phi}}^{(1)}) \cup ... \cup (\mathcal{L}_G^{(m)} \circ \mathcal{L}_{\boldsymbol{\phi}}^{(m)}) \cup ... \cup (\mathcal{L}_G^{(M)} \circ \mathcal{L}_{\boldsymbol{\phi}}^{(M)}) \tag{25}$$

where $\mathcal{L}_G^{(m)}$ is the *n*-gram model transducer and $\mathcal{L}_{\boldsymbol{\phi}}^{(m)}$ the interpolation weight transducer for the *m*th component. Note that the standard WFST *union* operation is performed at the complete sequence level, and is thus inappropriate for the *n*-gram level LM interpolation considered in Eqs. (2) and (15). Instead, a partial symbol sequence based, *n*-gram level *union* operation is required.

6.2. Log-linear model combination

Assuming compatible symbols are used for all transducers, a log-linear model combination may be efficiently implemented using a sequence of WFST *composition* operations between component *n*-gram model transducers after scaling of arc costs by their respective log-linear weights. This is given by

$$\mathcal{L} = (\mathcal{L}_G^{(1)} \times \lambda_1) \circ ...(\mathcal{L}_G^{(m)} \times \lambda_2) \circ ...(\mathcal{L}_G^{(M)} \times \lambda_M) \tag{26}$$

6.3. Multi-level model combination

As discussed in Section 2.2, in order to incorporate richer linguistic constraints, it is possible to train and combine LMs that model different linguistic unit sequences, for example, syllables and words (Liu et al., 2010). Linearly interpolated LMs built at the word and syllable level are intersected to yield a final combined multi-level LM. This LM leverages from both linear and log-linear forms of model combination and aims to achieve a good balance between generalization and discrimination. Using a WFST based representation, this form of hierarchical LM interpolation can be implemented by expanding a word level interpolated LM scored lattices into sub-word, e.g. the syllable level, via a *composition* with the *inverse* of the lexicon transducer which provides word to sub-word sequence mapping, before a final *composition* with a sub-word level interpolated LM. This effectively combines the linear and log-linear interpolation schemes given in Eqs. (25) and (26).

6.4. Interpolation weight set combination

As discussed in Section 5.3, when adapting LMs using context dependent interpolation, two sets of weights are available. These are estimated on the training data nPP estimation and test data recognition hypotheses respectively. The combination of these two sources of information can leverage the strength of both to improve LM adaptation. The associated combination schemes presented in Section 5.3 can also be efficiently implemented using WFSTs.

For the "MAP estimation and union" method, the final combined weight transducer for the *m*th component may be derived using

$$\mathcal{L}_{\boldsymbol{\phi}_m}^{\text{comb}} = \mathcal{L}_{\boldsymbol{\phi}_m}^{\text{MAP}} \cup \overline{\mathcal{L}}_{\boldsymbol{\phi}_m}^{\text{hier}} \tag{27}$$

where $\mathcal{L}_{\boldsymbol{\phi}_m}^{\text{MAP}}$ is the transducer of the nPP weights MAP adapted to the test set supervision using Eq. (22), and $\overline{\mathcal{L}}_{\boldsymbol{\phi}_m}^{\text{hier}}$ the weight transducer of contexts only observed in the adaptation supervision and estimated using the hierarchical smoothing of Eq. (18).

The "log-linear composition" scheme can be easily implemented via the WFST *composition* operation between the two interpolation weight transducers. This is given by,

$$\mathcal{L}_{\boldsymbol{\phi}_m}^{\text{comb}} = \mathcal{L}_{\boldsymbol{\phi}_m}^{\text{nPP}} \circ \mathcal{L}_{\boldsymbol{\phi}_m}^{\text{hier}} \tag{28}$$

where $\mathcal{L}_{\boldsymbol{\phi}_m}^{\text{nPP}}$ and $\mathcal{L}_{\boldsymbol{\phi}_m}^{\text{hier}}$ are the two transducers that represent the nPP prior estimated on the training data and the context dependent weights adapted to the test data hypotheses.

Finally, when using the "log-linear composition and union" based combination, the final combined weight set is derived as

$$\mathcal{L}_{\boldsymbol{\phi}_m}^{\text{comb}} = \left(\overline{\mathcal{L}}_{\boldsymbol{\phi}_m}^{\text{nPP}} \circ \mathcal{L}_{\boldsymbol{\phi}_m}^{\text{hier}}\right) \cup \mathcal{L}_{\boldsymbol{\phi}_m}^{\text{hier}} \tag{29}$$

where $\overline{\mathcal{L}}_{\boldsymbol{\phi}_m}^{\text{nPP}}$ is the weight transducer representing contexts that are only observed in the training data but not those in the adaptation supervision.

6.5. Decoding with context dependent LM interpolation

When using context dependent interpolation weights in decoding, there is a flexible choice between a static, off-line application, and dynamic, on-the-fly application of the weights. The static application is based on the WFST operations defined in Eq. (25) as previously discussed in Section 6. When interpolated language models using the nPP criterion are used, the interpolation weights are fixed. Hence, an off-line interpolation over all the n-grams in the component models may be performed. In order to compress the final network for efficiency, it is possible to use the conventional WFST *determinization* and *minimization* operations.

In contrast, during test-time LM adaptation, every broadcast show or snippet, for example, may have its own set of LM interpolation weights. When modelling a large number of contexts with distinct interpolation weights, the composition between component n-gram models and their weight transducers can lead to a significant expansion of the interpolated LM network. Many paths with unique LM scores will need to be kept distinct. Because the standard WFST network compression operations can only be performed statically in stages, a very large increase in the memory requirement during composition will occur. Hence, it is more efficient to dynamically perform the *composition*, *union* and *compression* operations in one single step on-the-fly. Component n-gram probabilities and interpolation weights will be applied for each context on request during decoding. Similar approaches have been previously shown to be effective for the composition between one single back-off n-gram LM and a lexicon transducer (Caseiro and Trancoso, 2006; Cheng et al., 2007; McDonough et al., 2007; Oonishi et al., 2009).

For the form of context dependent LM interpolation considered in this paper, the basic idea to only create a new path during search, if and only if it carries history context information that is different from others. The LM state associated with context history is *jointly* determined by component n-gram models and interpolation weights in the form of a context pair. Using this on-the-fly expansion algorithm, there are two major advantages. First, under the lattice constraint, no dead-end states (Caseiro and Trancoso, 2006), which have no path to the terminal state, will be created during expansion. Secondly, redundant paths representing unused lower order back-off distributions will be automatically filtered. The corresponding pseudo-code algorithm for an on-the-fly lattice expansion using context dependent LM interpolation is given below.

```
1:  for every node n_i in the network do
2:     initialize its expanded node list N'_i = {};
3:     initialize its expanded outbound arc list A'_i = {};
4:     initialize its LM state S_i = (null, null);
5:  end for
6:
7:     add n_0 to its expanded node list, N'_0 = {n_0};
8:     add all of n_0's outbound arcs to its expanded arc list, A'_0 = A_0;
9:
10:    Start depth first network traversal from the initial node n_0;
11:
12: for every node n_i being visited do
13:    for every expanded node n'_j ∈ N'_i of node n_i do
14:       for every outbound arc a_k from n_i do
15:          find the destination node n_k of arc a_k;
16:          find the LM state S'_j of expanded node n'_j;
17:          compute the interpolated LM probability P(n_k|S'_j);
18:          find longest context h_G in component LMs matching (S'_j, n_k);
```

find longest context h_{ϕ} in weight model matching (S'_j, n_k);
make context pair (h_G, h_{ϕ}) as a new LM state;
if $\exists$ node $n'_l \in N'_k$ representing LM state (h_G, h_{ϕ}) **then**
return the found node n'_l;
else
add a new node n'_l to N'_k to represent LM state (h_G, h_{ϕ});
end if
create a new arc a'_l from n'_i to n'_l;
assign LM score ln $P(n_k|S'_j)$ to a'_l;
copy other modeling info from a_k to a'_l;
add arc a'_l to the expanded outbound arc list A'_i for node n_i.
end for
end for
end for

Re-build new network with all expanded nodes and outbound arcs lists.

The above on-the-fly lattice expansion algorithm was implemented as an extension to the CU-HTK lattice processing tools. In practice, context dependent LM interpolation was found to only lead to a modest network size increase of between 20% and 120% compared to using standard context independent LM interpolation.

7. Experiments and results

In this section experimental results on a Mandarin Chinese broadcast speech transcription task are presented. First, two baseline LVCSR systems are described. Then the performance of various language models using the context dependent interpolation and adaptation schemes are evaluated. This is followed by experimental results on using the weight set combination methods proposed in Section 5.3. Finally, experimental results for the adaptation of a multi-level LM modelling both syllable and word sequences using context dependent interpolation are presented. These experiments were designed to investigate the following major topics:

- performance of the normalized perplexity metric based context dependent LM interpolation using the training data as presented in Section 5.2;
- performance of the context dependent LM adaptation techniques proposed in Section 5.1 using various forms of weight smoothing priors presented in Sections 5.1 and 5.3;
- performance of the MBR based discriminative LM adaptation scheme presented in Section 3.

7.1. Baseline system description

The 2006 CU-HTK Mandarin Chinese LVCSR system developed for the DARPA GALE phase I evaluation was initially used to evaluate LMs employing the various interpolation and adaptation techniques proposed in Section 5. The overall structure of the system was similar to that described in Sinha et al. (2006). It comprises an initial lattice generation stage using a 58k word list, interpolated 4-gram word based back-off LM, and adapted MPE (Povey and Woodland, 2002) acoustic models trained on HLDA (Kumar, 1997; Liu et al., 2003) projected PLP (Hermansky, 1990; Woodland et al., 1996) features augmented with pitch parameters. A total of 942 h of audio data containing mixed broadcast news (BN) and broadcast conversation (BC) speech genre were used for acoustic model training. A total of 1.0G words from 10 text sources were used in baseline LM training. Information on corpus size, cut-off settings and smoothing schemes for component LMs are given in Table 1. For data sources that are closer in genre to the test data, the lowest cut-offs and modified KN smoothing were used. These include the two acoustic transcriptions sources, bcm and bnm, and additional data collected from major TV channels or media such as CCTV, VOA and Phoenix TV. For example, a cut-off of "111" were used for the bnm and bcm sources, as are shown in the 3rd column of the first two lines in Table 1. This setting implies that there has to be at least one occurrence of any bigram, trigram or fourgram if any of them were to be retained. For the two largest corpora of newswire genre, giga-xin and giga-cna, more aggressive cut-offs and Good Turing (GT) discounting were used. As discussed in Section 5.2, these conscious decisions are often made in state-of-the-art LVCSR systems when certain text sources are known to

Table 1
Text source, 2/3/4-gram cut-off settings, smoothing scheme used in training, model size and global ML weights tuned using test set PP (bn06 + bc05), training data PP and nPP scores for component language models of the 2006 CU-HTK Mandarin Chinese LVCSR system.

Comp LM	Text (M)	Train config	Model size (M)			Global weight tuning		
			2g	3g	4g	Base	PP	nPP
bcm	4.83	111,kn	1.19	3.06	3.78	0.233	0.005	0.143
bnm	3.78	111,kn	1.07	2.45	2.91	0.133	0.007	0.173
giga-xin	277.6	123,gt	19.3	26.1	10.4	0.139	0.243	0.108
giga-cna	496.7	123,gt	24.9	37.1	12.2	0.173	0.458	0.082
phoenix	76.89	112,kn	11.5	40.1	8.34	0.103	0.113	0.123
voarfabbc	30.28	112,kn	2.99	9.24	1.97	0.103	0.027	0.073
cctvcnr	26.81	112,kn	5.16	15.2	2.74	0.040	0.039	0.084
tdt4	1.76	112,kn	0.71	1.35	0.09	0.025	0.006	0.072
papersjing	83.73	122,kn	9.43	10.2	11.3	0.029	0.092	0.093
ntdtv	12.49	122,kn	2.27	1.27	1.23	0.032	0.011	0.050

be more useful for the target domain. Three Mandarin broadcast speech evaluation sets were used: bn06 of 3.4 h BN data, bc05 of 2.5 h of BC data and the 1.8 h GALE 2006 evaluation set eval06 containing a mixture of BN and BC data.

The 2008 CU-HTK Mandarin Chinese LVCSR system developed for the DARPA GALE phase III evaluation was then used to evaluate performance of history context dependently adapted multi-level LMs in the final part of this section. Compared with the first 2006 baseline system described above, additional data was used in model training. The acoustic models were trained on 1673 h of speech. A total of 4.3 billion characters from 27 text sources were used in LM training. These account for 2.8 billion words after a longest first based character to word segmentation. A larger 63k word list consisting a total of 52k multiple character Chinese words, 6k single character Chinese words and 5k frequent English words was used. Information on corpus size, cut-off settings, smoothing schemes and component weights for the top 10 heavily weighted text sources are given in Table 2.

The CU-HTK Mandarin Chinese LVCSR system uses a multi-pass recognition decoding framework (Sinha et al., 2006). Unadapted acoustic models and the baseline word level 4-gram LM were used to the first recognition pass "P1" to generate initial hypotheses for subsequent acoustic model adaptation. Adapted acoustic models and the baseline 4-gram LM were then used in the following "P2" lattice generation stage. This was followed by a "P3" lattice rescoring stage using re-adapted acoustic models. Three GALE Mandarin broadcast speech development sets of mixed BN and BC genre were used: 2.6 h dev07, 1 h dev08 and 2.6 h p2ns. Manual audio segmentation was used. The word level baseline LM component weights were perplexity tuned on dev07, dev08, bn06 and bc05.

Table 2
Text source size, cut-off settings, smoothing scheme used and interpolation weights for top 10 heavily weighted text sources for component language models of the 2008 CU-HTK Mandarin Chinese LVCSR system. Cut-off settings for 5-gram and 6-gram character level LMs are shown in brackets.

Comp LM	#Char (M)	#Word (M)	Train config	Interpolation weight
bcm	14.26	9.21	kn/111(11)	0.260058
bnm	12.29	7.41	kn/111(11)	0.147834
gigaxin	483.65	362.74	kn/112(22)	0.132539
phoenix	144.57	91.38	kn/112(22)	0.107920
gigacna	891.13	604.98	gt/123(33)	0.072665
voarfa	63.54	35.31	kn/112(22)	0.072299
ibmsina2	382.34	253.59	kn/112(22)	0.055601
bbndata	301.39	186.3	kn/112(22)	0.046213
galeweb	556.41	390.8	kn/122(22)	0.045918
agilece	336.78	204.5	kn/112(22)	0.031497

Table 3
PP and lattice rescoring 1-best CER% performance of interpolated LMs on bn06, bc05 and eval06. Global interpolation weights of the baseline LM tuned on the reference of combined bn06 + bc05 set. Equal weights initialization for all models.

LM	History context	nPP	Reference perplexity			CER%		
			bn06	bc05	eval06	bn06	bc05	eval06
Base	–	84.0	194.6	227.1	231.7	8.4	19.0	19.1
PP	–	130.8	224.5	403.8	360.7	8.6	20.7	19.9
nPP	–	82.0	197.6	239.9	235.3	8.3	19.1	19.1
	1g	69.7	193.3	224.1	221.9	8.1	19.1	19.1
	3g	51.9	179.3	213.1	215.2	8.1	19.0	18.7

7.2. Performance of interpolated language models

The PP and nPP scores for various interpolated LMs are presented in Table 3. The first line shows the performance of a baseline interpolated LM using global, context independent weights that were perplexity tuned on the combined bn06 + bc05. This is the standard form of model interpolation for current ASR systems. These weights are in the 7th column of Table 1. The second line shows the performance of weights tuned using the training data perplexity. Compared to the baseline system, there was a large degradation of 30-176 PP points on the all three test sets. Similarly there is a large error rate increase of 0.2–1.7% absolute. This is due to the corpus size bias previously discussed in Section 5.2. Such a bias further manifests itself in the global weights given in the 8th column of Table 1. The largest two corpora, giga-cna (0.46) and giga-xin (0.24) were heavily weighted.

Using the nPP metric in context independent weight estimation, this bias was greatly reduced, as given in the third line of Table 3. The corresponding global weights are in the last column of Table 1. Large sized corpora no longer dominate the weight assignment. As discussed in Sections 1 and 5.2, the weights are determined by a combination of global factors, such as source of collection, epoch and genre, and local factors including modeling resolution, generalization, topics and styles. During the nPP estimation, if a particular component model is both under-fitting to its own training data and generalizing poorly to other sources, its weight is likely to be low. For example, the biggest newswire source, giga-cna, is of Taiwanese origin and different in style from other broadcaster sources, trained using aggressive cut-offs and simple GT discounting, is now weighted by 0.082. In contrast, the acoustic transcription source bnm, collected from major mainland Chinese broadcasters, is similar in genre, topics and style to most of other data sources in the table, and trained with minimum cut-offs and more robust KN smoothing, is weighted by 0.17. It is also interesting to note that if minimum cut-offs and KN smoothing were used to build all source specific models and no conscious bias is introduced, a much smoother weight distribution can be obtained. For example, the weight assigned to bnm is decreased to 0.13. This is expected as an improved modelling resolution and smoothing scheme would help certain sources to be more useful during nPP estimation, and therefore be more competitive against other sources. Using the "nPP" system with context independent LM interpolation weights, perplexity and CER performance comparable to the baseline LM was obtained. 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. The performance of two context dependent systems are shown in the last two lines of Table 3. Due to memory constraints in weight estimation, the three word history based "3g" nPP model was built by extracting word level contexts from the single word history based "1g" nPP model after being pruned at 1.0e−9. A total of 58k single word, 5.2M two word and 1.9M three word history contexts were retained to have their own weights. These were trained using the nPP estimation with hierarchical weight smoothing given in Eq. (21). Using 1-gram word level weights, there are more than 10 points of PP improvement on eval06, and an absolute CER reduction of 0.3% over the baseline on bn06. Increasing the context span to 3-gram word based history gave the best PP and CER performance. Compared with the baseline LM, 14–16 points of PP improvements (7% relative) were obtained over all test sets. This system also outperformed the baseline LM on bn06 and eval06 with a statistically significant[2] CER

[2] For all results presented in this paper, matched pairs sentence-segment word error (MAPSSWE) based statistical significance test was performed at a significance level $\alpha = 0.05$.

Table 4
CER and PP performance of ML adapted 4-gram LMs on bn06, bc05 and eval06. Equal weight initialization for all adapted models. Context dependent adaptation uses a hierarchical smoothing prior.

LM adapt	History context	Reference perplexity			CER%		
		bn06	bc05	eval06	bn06	bc05	eval06
Base-adapt	–	150.0	201.1	200.9	8.1	18.9	18.8
ML	1g	138.5	188.3	187.7	8.1	18.9	18.7
	3g	132.9	176.0	184.2	8.0	18.8	18.7

improvements of 0.3–0.4% absolute. Overall, there is a consistent and strong correlation between training data nPP and test set PP scores in Table 3.

7.3. Performance of adapted language models

The performance of adapted LMs using context dependent interpolation weights was evaluated next. LM adaptation was performed at the audio document level. Both bn06 and bc05 consist of a number of 0.5 h long broadcast shows, while eval06 is made up of broadcast snippets of five minutes on average. Equal weight initialization was used. The smoothing constant setting $\tau = 2.5$ were used in the MAP adaptation given in Eq. (18). The MBR smoothing constant D is set as $D = E \times N_{\mathcal{W}}$, where $N_{\mathcal{W}}$ is the number of speech segments in the supervision data, and $E = 50$. A total of 8 iterations of weight re-estimation were performed. The 4-gram lattice 1-best output generated by the baseline LM with global, context independent weights (first line of Table 3) was used as the adaptation supervision. The top 1000 hypotheses were extracted for MBR adaptation. During decoding component language models were combined on-the-fly using adapted interpolation weights and the WFST based on-the-fly lattice expansion algorithm described in Section 6. As discussed in Section 5, an important issue during MAP estimation of context dependent weights is the form of the smoothing prior. Experiments in this section only consider using the hierarchical weight smoothing given in Eq. (18) to serve as baseline context dependent adaptation configuration. In practice this was found to consistently outperform a global, context independent prior.

The performance of ML adaptation is shown in Table 4. Using global, context independent PP based adaptation, there are 26 to 45 points of PP improvements (12–23% relative) for all sets over the unadapted baseline system in the first line of Table 3. Absolute CER gains of 0.3% on bn06 and eval06 were obtained. Using context dependent adaptation, a further PP reduction of 13 points (8% relative) was obtained by the word level 1-gram weights. However, the CER gains were marginal. Using longer 3-gram word context based weights gave the best adaptation performance. On average for each audio document, approximately 0.9k single word, 2.2k two word and 2.5k three word history contexts were adapted to have their own weights. This system gave a PP reduction of 17–25 points (8–12% relative) on all test sets against the baseline context free adaptation configuration, but only transformed into marginal CER gains of 0.1% absolute. These results suggest a weak correlation between PP and error rate[3]. ML based language model adaptation may improve PP on the common contexts observed in both the supervision and reference, but is not necessarily helpful in generalization and discrimination. Hence, it would be interesting to evaluate the performance of MBR adaptation.

The results of MBR adaptation are shown in Table 5. Improved performance was obtained using MBR estimated 3-gram word level context based weights, as is shown in the bottom line of the table. Consistent CER improvements of 0.1–0.3% absolute were obtained over the ML adapted context independent baseline and the context dependent system. In total, this system gave statistically significant CER reductions of 0.4% on bn06, 0.4% on bc05 and 0.5% on eval06 over the system without LM adaptation shown in the first line of Table 3.

[3] It is expected than when LM adaptation uses information from the recognition hypotheses, the adapted LM no longer provides an "independent" prior information when combined with acoustic models. Hence, the improvement in perplexity may not necessarily lead to lower error rates.

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

LM adapt	History context	CER%		
		bn06	bc05	eval06
Base-adapt	–	8.1	18.9	18.8
ML	3g	8.0	18.8	18.7
MBR	3g	8.0	18.6	18.6

Table 6
PP performance of adapted 4-gram LMs on bn06, bc05 and eval06.

History context		Wgt comb	PP (reference)		
Prior	Adapt		bn06	bc05	eval06
Base-adapt		–	150	201	201
–	3g	–	133	176	184
3g		map	128	176	179
	3g	map + union	120	168	171
		log	126	180	180
		log + union	118	166	169

7.4. Combined use of interpolation and adaptation weights

As discussed in Section 5.3, when adapting LMs using context dependent interpolation, two sets of weights are available. These are obtained from the training data nPP estimation and test data adaptation respectively. The trade-off between using domain independent, longer span context weights estimated on the training data, and in-domain, shorter context weights adapted using recognition hypotheses is an important issue. In this section experiments are conducted to evaluate a range of weight combination methods proposed in Section 5.3. The PP and CER performance of various language models are shown in Tables 6 and 7. The first line in both tables is the baseline adapted system shown in the first line of Table 4 using context independent weights. The performance of the three word history context dependent, perplexity adapted LM without using training set information is also shown in the second line of Tables 6 and 7. The CER and perplexity performance of this system was also previously shown in the bottom line of Table 4.

The values of perplexity and CER performance of combining training set nPP and test adapted weights using the methods proposed in Section 5.3 are shown in the last four lines of Tables 6 and 7 respectively. The MAP estimation and log-linear composition approaches gave similar performance when only adapting weights for contexts observed in the training data. Using the MAP estimation with union, a further PP reduction was obtained but there were no further CER gains. Note that the map + union system did not outperform the second system in the two tables where a simpler hierarchical smoothing was applied to all contexts and no combination with training set nPP weights was performed. This may be caused by the mismatch of the nPP priors used in Eq. (23) against the target domain during adaptation. As discussed in Section 5.3, such a mismatch may be partially retained during MAP estimation. The log-linear composition

Table 7
CER performance of adapted 4-g LMs on bn06, bc05 and eval06.

History context		Wgt comb	CER%		
Prior	Adapt		bn06	bc05	eval06
Base-adapt		–	8.1	18.9	18.8
–	3g		8.0	18.8	18.7
3g		map	8.1	18.9	18.7
	3g	map + union	8.0	18.9	18.7
		log	8.0	19.0	18.7
		log + union	7.9	18.8	18.5

Table 8
The hit rate statistics measured on the reference transcription of bn06, bc05 and eval06 for word level context dependent interpolation weight vectors of varying length. These three sets were obtained from the nPP prior, adapting to the recognition hypotheses, or the combined weight set derived using option D respectively.

History context	Reference hit rate% (1g/2g/3g)		
	bn06	bc05	eval06
Train	16.9/55.4/27.7	16.1/55.2/28.8	19.2/53.5/27.3
Test	15.7/10.0/69.6	26.7/19.4/49.8	25.4/13.3/53.7
Comb	17.5/12.5/70.1	28.4/21.7/50.4	28.4/17.6/54.4

Table 9
CER performance of MBR adapted 4-gram LMs on bn06, bc05 and eval06.

History context		Wgt comb	CER%		
Prior	Adapt		bn06	bc05	eval06
Base-adapt		–	8.1	18.9	18.8
3g	3g-ML	log + union	7.9	18.8	18.5
	3g-MBR		7.9	18.6	18.5

and union approach gave the best performance among the four. More than 30 points of PP reduction (17–22% relative) and 0.1–0.3% absolute CER reduction were obtained over the "base-adapt" baseline system using context free weights (1st line of Tables 6 and 7). The associated hit rate statistics on the reference transcription for this system on various test sets are shown in the bottom line of Table 8. The first line of the table shows the history context hit rates using the nPP prior model alone. Compared to the hit rates using the weights obtained from test set adaptation alone shown in the second line, the combined weight set derived from the union operation of option D consistently improved the context coverage on the reference transcriptions for all three sets. In addition to the perplexity and CER results Tables 6 and 7, these statistics further suggest sufficient coverage of contexts in test set supervision is important when adapting LMs using context dependent interpolation.

As previously shown in Table 5, MBR adaptation gave improved discrimination and CER performance over MAP adaptation. Hence, it is now interesting to investigate the performance of MBR discriminative adaptation with a nPP weight prior, as is shown in Table 9. The log-linear composition and union based approach (option D) to combine with the "3g" nPP weights of Table 3 gave a further small CER improvement of 0.2% on bc05 against a comparable ML adaptation configuration. The total CER gains over the unadapted baseline system (also shown in first line of Table 3) are 0.5% (6% relative) on bn06, 0.4% on bc05 and 0.6% on eval06, all being statistically significant.

7.5. Multi-level LM combination and adaptation

As discussed in Section 6, in order to incorporate richer linguistic constraints, it is possible to train and combine LMs that model different unit sequences, for example, syllables and words (Hieronymus et al., 2009; Liu et al., 2010). Context dependent interpolation was used to build LMs at the word and syllable level before intersected to yield a final combined multi-level LM. This LM leverages both linear and log-linear forms of model combination and aims to achieve a good balance between generalization and discrimination. The use of context dependent weights within each layer of the modelling hierarchy also allows a multi-level LM to be adapted using the methods proposed in this paper. The performance of this adapted multi-level LM was evaluated on the 2008 CU-HTK Mandarin Chinese LVCSR system described in Section 7.1.

The confusion network (CN) decoding performance of the baseline word level LM is shown in the first line of Table 10. In order to incorporate additional sub-word level constraints in LMs, syllable level LMs can be constructed and combined with word level LMs. One issue with this method is that syllable segmented and labelled Chinese texts are expensive to produce and generally unavailable in large quantities. The difficulty arises from the fact that many Chinese characters have multiple pronunciations. Since Chinese characters are syllabic in nature, an alternative is to use character level LMs as an indirect way of modeling syllable sequences (Hieronymus et al., 2009; Liu

Table 10
Performance of language models on dev07, dev08 and p2ns. "'○" denotes the WFST composition operation.

P2 system	LM adapt	CER%		
		dev07	dev08	p2ns
w.4g	–	9.7	9.6	9.6
c.6g	–	10.9	10.0	10.3
w.4g ○ c.6g	–	9.5	9.1	9.3
w.4g	CI	9.6	9.3	9.4
w.4g	CD	9.5	9.2	9.3
w.4g ○ c.6g	CD	9.4	8.9	9.1

Table 11
CN performance of P3 acoustic rescoring of P2 lattices generated by various language models on dev07, dev08 and p2ns.

P3 system	LM adapt	CER%		
		dev07	dev08	p2ns
w.4g	–	9.3	8.7	9.1
w.4g	CI	9.1	8.6	9.1
w.4g	CD	9.0	8.5	8.8
w.4g ○ c.6g	CD	8.8	8.4	8.6

et al., 2010). 6-gram character level LMs were built and linearly interpolated. Their cut-off settings are shown in brackets of Table 2. On average the word based system produces approximately 1.5 characters per word. Hence, a 6-gram character level LM has a comparable context span to word level 4-gram LMs. The performance of this system is shown in the second line of Table 10. As expected, with a stronger constraint, the word level 4-gram baseline significantly outperformed the character 6-gram LM alone by 0.4–1.2% absolute. When combining syllable and word constraints using an equal weighted log-linear interpolation of Eq. (3) and the WFST representation of Eq. (26), consistent performance improvements were obtained over the word level baseline. This is shown in the 3rd line of Table 10. It gave statistically significant CER reductions of 0.5% and 0.3% on dev08 and p2ns respectively.

The second section of Table 10 shows the performance of three adapted LMs using the WFST representation in Eq. (25). The 1-best output from the un-adapted word level baseline system was used as the supervision in perplexity based LM adaptation. Standard LM adaptation using context independent interpolation weights gave CER reductions of 0.1–0.3% absolute across three test sets (4th line of Table 10). Using the context dependent adaptation of Eq. (15) with a hierarchical smoothing prior, and intersected with a high resolution training data nPP prior (the log-linear combination and union approach discussed in Section 5.3), a further CER improvement of 0.1% absolute was obtained for all test sets (5th line of Table 10). Adapting both the word and character level LMs using context dependent weights before a final log-linear combination gave the best performance in the table. Absolute CER reductions of 0.4% and 0.3% on dev08 and p2ns were obtained over the baseline word level LM adapted using context independent interpolation. The total performance improvements over the unadapted word level baseline are 0.3% on dev07, 0.7% dev08 (7.3% relative) and 0.5% on p2ns (5.2% relative) respectively, all being statistically significant.

Table 10 shows the performance of multi-level combined and adapted LMs at the lattice generation stage. Now it's interesting to examine if the performance improvements can be maintained at a later pass of the recognition system where re-adapted acoustic models are used to rescoring lattices generated by various LMs in Table 10. These are shown in Table 11. The performance gains from the adapted multi-level combined LM (last line of Table 11) over the word level baseline (first line of Table 11) were largely maintained. Statistically significant CER reductions of 0.3–0.5% absolute were obtained over all test sets, in particular, 0.5% absolute (5.5% relative) for dev07 and p2ns.

7.6. Discussion

A range of experiments were conducted to evaluate the performance of context dependent LM interpolation and adaptation techniques presented in this paper. Experimental results lead to the following findings:

- it is possible to build a domain or task neutral LM on the training data using context dependent interpolation weights that are estimated using the normalized perplexity criterion presented in Section 5.2;
- the context dependent LM adaptation techniques proposed in Sections 5.1 and 5.3 are useful to improve the performance of state-of-the-art LVCSR systems;
- the MBR based discriminative LM adaptation scheme presented in Section 3 gave further improvements over conventional maximum likelihood based approaches.

8. Conclusion

A context dependent form of language model interpolation and adaptation was investigated in this paper. A novel LM interpolation technique using normalized perplexity was proposed to robustly estimate context dependent language model interpolation weights on the training data. MAP estimation of back-off weights was also used to address the data sparsity problem. Several forms of smoothing priors were proposed. A range of schemes to use weight estimates from both training data and test data hypothesis were proposed to improve robustness in LM adaptation. A WFST based on-the-fly decoding algorithm for context dependent LM interpolation is also presented. Experimental results on a state-of-the-art Mandarin Chinese broadcast speech transcription task suggest that context dependent language model interpolation and adaptation may be useful for speech recognition. Future research will focus on improving robustness of LM adaptation. A continuous representation (Bengio and Ducharme, 2003; Schwenk, 2007) of the history weighting function will also be investigated.

References

Bahl, L., et al., 1995. The IBM large vocabulary continuous speech recognition system for the ARPA NAB news task. In: Proceedings of the ARPA Workshop on Spoken Language Technology, pp. 121–126.

Bellegarda, J., 2000. Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE 88 (August (8)), 1279–1296.

Bengio, Y., Ducharme, R., 2003. A neural probabilistic language model. Advances in Neural Information Processing Systems 3, 1137–1155.

Blei, D., Ng, A., Jordan, M., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3.

Brants, T., 2005. Test data likelihood for PLSA models. Information Retrieval 8 (2), 181–196.

Bulyko, I., Ostendorf, M., Stolcke, A., 2012. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In: Proceedings of the HLT'03, Edmonton.

Bulyko, I., Matsoukas, S., Schwartz, R., Nguyen, L., Makhoul, J., 2007. Language model adaptation in machine translation from speech. In: Proceedings of the ICASSP'07, Hawaii.

Caseiro, D.A., Trancoso, I., 2006. A specialized on-the-fly algorithm for lexicon and language model composition. IEEE Transactions on Audio, Speech, and Language Processing 14 (4), 1281–1291.

Chen, S.F., Goodman, J.T., 1999. An empirical study of smoothing techniques for language modeling. Computer Speech and Language 13 (4), 359–394.

Chen, L., Gauvain, J.-L., Lamel, L., Adda, G., Adda, M., 2001. Language model adaptation for broadcast news transcription. In: Proceedings of the ISCA ITRW'01, Paris.

Cheng, O., Dines, J., Doss, M.M., 2007. A generalized dynamic composition algorithm of weighted finite state transducers for large vocabulary speech recognition. In: Proceedings of the ICASSP'07, Hawaii.

Chien, J.T., Wu, M.S., Wu, C.S., 2005. Bayesian learning for latent semantic analysis. In: Proceedings of the Interspeech'05, Lisbon.

Clarkson, P., Robinson, A., 1997. Language model adaptation using mixtures and an exponentially decaying cache. In: Proceedings of the ICASSP1997, Munich, pp. 799–802.

Darroch, J., Ratcliff, D., 1972. Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics 43 (5), 1470–1480.

Doumpiotis, V., Byrne, W., 2005. Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition. Speech Communication 2, 142–160.

Federico, M., 1999. Efficient language model adaptation through MDI estimation. In: Proceedings of the EuroSpeech'99, Budapest.

Federico, M., 2003. Language model adaptation through topic decomposition and MDI estimation. In: Proceedings of the IEEE ICASSP2003, Hong Kong.

Gildea, D., Hofmann, T., 1999. Topic-based language models using EM. In: Proceedings of the Eurospeech'99, Budapest.

Gopalakrishnan, P.S., Kanevsky, D., Nádas, A., Nahamoo, D., 1991. An inequality for rational functions with applications to some statistical estimation problems. IEEE Transactions on Information Theory 37 (January (1)).

Hermansky, H., 1990. Perceptual linear prediction (PLP) of speech. Journal of the Acoustic Society of America 87 (4), 1738–1752.
Hinton, G., 2002. Training products of experts by minimizing contrastive divergence. Neural Computation 14, 1771–1800.
Hieronymus, J.L., Liu, X., Gales, M.J.F., Woodland, P.C., 2009. Exploiting Chinese character models to improve speech recognition performance. In: Proceedings of the Interspeech'09, Brighton.
Hsu, B., 2007. Generalized linear interpolation of language models. In: Proceedings of the IEEE ASRU'07, Kyoto.
Iyer, R., Ostendorf, M., Rohlicek, R., 1994. An improved language model using a mixture of Markov components. In: Proceedings of the ARPA Workshop on Human Language Technology, Plainsboro, NJ, pp. 82–87.
Iyer, R., Ostendorf, M., 1999. Modeling long distance dependence in language: topic mixtures vs. dynamic cache models. IEEE Transactions on Speech and Audio Processing 7 (January (1)), 30–39.
Jelinek, F., Mercer, R., 1980. Interpolated estimation of Markov source parameters from sparse data. In: Gelsema, E.S., Kanal, L.N. (Eds.), Pattern Recognition in Practice. North Holland, Amsterdam, pp. 381–402.
Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press, Cambridge, Massachusetts.
Katz, S.M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 35 (3), 400–401.
Kaiser, J., Horvat, B., Kacic, Z., 2012. A novel loss function for the overall risk-criterion based discriminative training of HMM models. In: Proceedings of the ICSLP'00, Beijing.
Kneser, R., Steinbiss, V., 1993. On the dynamic adaptation of stochastic LM. In: Proceedings of the ICASSP1993, vol. 2, Minneapolis, pp. 586–589.
Kneser, R., Peters, J., 1997. Semantic clustering for adaptive language modeling. In: Proceedings of the ICASSP1997, vol. 2, Munich, pp. 779–782.
Kumar, N., 1997. Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition. PhD Thesis. John Hopkins University, Baltimore.
Liu, X., Gales, M.J.F., Woodland, P.C., 2003. Automatic complexity control for HLDA systems. In: Proceedings of the IEEE ICASSP2003, vol. 1, Hong Kong, pp. 132–135.
Liu, X., Byrne, W.J., Gales, M.J.F., Woodland, P.C., et al., 2007. Discriminative language model adaptation for Mandarin broadcast speech transcription and translation. In: Proceedings of the IEEE ASRU'07, Kyoto.
Liu, X., Gales, M.J.F., Woodland, P.C., 2008. Context dependent language model adaptation. In: Proceedings of the Interspeech'08, Brisbane.
Liu, X., Gales, M.J.F., Woodland, P.C., 2009. Use of contexts in language model interpolation and adaptation. In: Proceedings of the Interspeech'09, Brighton.
Liu, X., Gales, M.J.F., Hieronymus, J.L., Woodland, P.C., 2010. Language model combination and adaptation using weighted finite state transducers. In: Proceedings of the IEEE ICASSP2010, Dallas.
McDonough, J., Stoimenov, E., Klakow, D., 2007. An algorithm for fast composition of weighted finite-state transducers. In: Proceedings of the ASRU'07, Kyoto.
Mrva, D., Woodland, P.C., 2004. A PLSA-based language model for conversational telephone speech. In: Proceedings of the InterSpeech'04, Jeju.
Mrva, D., Woodland, P.C., 2006. Unsupervised language model adaptation for Mandarin broadcast conversation transcription. In: Proceedings of the ICSLP'06, Pittsburgh.
Mohri, M., 1997. Finite-state transducers in language and speech processing. Computational Linguistics 23 (2), 269–311.
Mohri, M., Riley, M., 1998. Network optimizations for large vocabulary speech recognition. Speech Communication 25 (3).
Mohri, M., Riley, M., Hindle, D., Ljolje, A., Pereira, F., 1998. Full expansion of context-dependent networks in large vocabulary speech recognition. In: Proceedings of the ICASSP'98, Seattle.
Mohri, M., Pereira, F., Riley, M., 2000. Weighted finite-state transducers in speech recognition. In: Proceedings of the ASR2000, Paris.
Normandin, Y., 1991. Hidden Markov models maximum mutual information estimation and the speech recognition problem. PhD Thesis. McGill University, Canada.
Och, F.J., Ney, H., 2012. Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the ACL02', Philadelphia, pp. 295–302.
Oonishi, T., Dixon, P., Iwano, K., Furui, S., 2009. Generalization of specialized on-the-fly composition. In: Proceedings of the ICASSP'09, Taipei.
Della Pietra, S., Della Pietra, V., Mercer, R.L., Roukos, S., 1992. Adaptive language modeling using minimum discriminant estimation. In: Proceedings of the ICASSP'92, San Francisco, pp. 1633–1636.
Povey, D., Woodland, P.C., 2002. Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of the ICASSP'02, Orlando.
Rosenfeld, R., 1996. A maximum entropy approach to adaptive statistical language modeling. Computer Speech and Language 10, 187–228.
Rosenfeld, R., 2000. Two decades of statistical language modeling: where do we go from here? Proceedings of the IEEE 88 (August (8)).
Rosenfeld, R., Chen, S.F., Zhu, X., 2001. Whole-sentence exponential language models: a vehicle for linguistic-statistical integration. Computers Speech and Language 15 (1), 2001.
Schwenk, H., 2007. Continuous space language models. Computer Speech and Language 21 (July (3)), 492–518.
Seymore, K., Rosenfeld, R., 1997. Using story topics for language model adaptation. In: Proceedings of the Eurospeech'97, Rhodes.
Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Proceedings of the Proceedings of the ICSLP'02, Denver.
Sinha, R., Gales, M.J.F., Kim, D.Y., Liu, X., Sim, K.C., Woodland, P.C., 2006. The CU-HTK Mandarin broadcast news transcription system. In: Proceedings of the ICASSP'06, Toulouse.
Tam, Y.C., Schultz, T., 2005. Dynamic language model adaptation using variational Bayes inference. In: Proceedings of the Interspeech'05, Lisbon.
Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J., 1996. The development of the 1996 HTK broadcast news transcription system. In: Proceedings of the DARPA Speech Recognition Workshop, Arden House, NY, USA, pp. 73–78.
Young, S., et al., 2009. The HTK Book Version 3.4.1.