Interpolated Dirichlet Class Language Model for Speech Recognition Incorporating Long-distance N-grams

Md. Akmal Haidar and Douglas O'Shaughnessy
INRS-EMT, University of Quebec
6900-800 De la Gauchetier Ouest, H5A 1K6, Montreal (Quebec), Canada
haidar@emt.inrs.ca, dougo@emt.inrs.ca

Abstract

We propose a language modeling (LM) approach incorporating interpolated distanced n-grams in a Dirichlet class language model (DCLM) (Chien and Chueh, 2011) for speech recognition. The DCLM relaxes the bag-of-words assumption and documents topic extraction of latent Dirichlet allocation (LDA). The latent variable of DCLM reflects the class information of an n-gram event rather than the topic in LDA. The DCLM model uses default background n-grams where class information is extracted from the $(n\text{-}1)$ history words through Dirichlet distribution in calculating n-gram probabilities. The model does not capture the long-range information from outside of the n-gram window that can improve the language modeling performance. In this paper, we present an interpolated DCLM (IDCLM) by using different distanced n-grams. Here, the class information is exploited from $(n\text{-}1)$ history words through the Dirichlet distribution using interpolated distanced n-grams. A variational Bayesian procedure is introduced to estimate the IDCLM parameters. We carried out experiments on a continuous speech recognition (CSR) task using the Wall Street Journal (WSJ) corpus. The proposed approach shows significant perplexity and word error rate (WER) reductions over the other approach.

1 Introduction

Statistical n-gram LMs have been successfully used for speech recognition and many other applications. They suffer from insufficiencies of training data and long-distance information, which limit the model generalization (Chien, 2006). The data sparseness problem is usually solved by backoff smoothing using lower-order language models (Katz, 1987; Kneser and Ney, 1995). The class-based language model was investigated where the class n-grams were calculated by considering the generation of concatenated classes rather than words (Brown et al., 1992). By incorporating the multidimensional word classes and considering the classes from various positions of left and right contextual information (Bai et al., 1998), the class n-gram can be improved (Yamamoto et al., 2003). A neural network language model (NNLM) was trained by linearly projecting the history words of an n-gram event into a continuous space (Bengio et al., 2003; Schwenk, 2007). Later, a recurrent neural network-based LM was investigated that shows better results than NNLM (Mikolov et al., 2010; Mikolov et al., 2011). Unsupervised class-based language models such as Random Forest LM (Xu and Jelinek, 2007), Model M (Chen, 2008) have been investigated that outperform a word-based LM. However, the long-distance information is captured by using a cache-based LM that takes advantage of the fact that a word observed earlier in a document could occur again. This helps to increase the probability of the seen words when predicting the next word (Kuhn and Mori, 1990).

To compensate for the weakness of the n-gram models, latent topic analysis has been used broadly. Several techniques such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Bellegarda, 2000), probabilistic LSA (PLSA) (Hofmann, 1999; Gildea and Hofmann, 1999), and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have been studied to extract the latent semantic information from a training

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1793–1802, Dublin, Ireland, August 23-29 2014.

corpus. The LSA, PLSA and LDA models have been used successfully in recent research work for LM adaptation (Bellegarda, 2000; Gildea and Hofmann, 1999; Mrva and Woodland, 2004; Tam and Schultz, 2005; Tam and Schultz, 2006; Haidar and O'Shaughnessy, 2011; Haidar and O'Shaughnessy, 2012b; Haidar and O'Shaughnessy, 2012a). Even so, the extracted topic information is not directly useful for speech recognition, where the latent topic of n-gram events should be of concern. In (chien and Chueh, 2008), a latent Dirichlet language model (LDLM) was proposed where the latent topic information was exploited from (n-1) history words through the Dirichlet distribution in calculating the n-gram probabilities. A topic cache language model was proposed where the topic information was obtained from long-distance history through multinomial distributions (Chueh and Chien, 2010). Topic-dependent-class-based n-gram LM was proposed where the LSA method was used to reveal latent topic information from noun-noun relations (Naptali et al., 2012). In (Bassiou and Kotropoulos, 2010), a PLSA technique enhanced with long-distance bigrams was used to incorporate the long-term word dependencies in determining word clusters. This technique was used in (Haidar and O'Shaughnessy, 2013b) and (Haidar and O'Shaughnessy, 2013a) for the PLSA and LDLM models respectively where the long-distance information was captured by using interpolated distanced n-grams and their parameters were estimated by using an expectation maximization (EM) procedure (Dempster et al., 1977). In (Chien and Chueh, 2011), the DCLM model was proposed to tackle the data sparseness and to extract the large-span information for the n-gram model. In this model, the topic structure in LDA is assumed to derive the hidden classes of histories in calculating the language model. A Bayesian class-based language model was presented where a variational Bayes-EM procedure was used to compute the model parameters. Also, a cache DCLM model was proposed to capture the long-distance information beyond the n-gram window. However, in the DCLM model (Chien and Chueh, 2011), the class information of the history words was obtained from the n-gram events of the corpus. Here, the long-range information outside the n-gram window is not captured. In this paper, we present an IDCLM model to capture the long-range information in the DCLM using the interpolated distanced n-grams. The n-gram probabilities of the proposed IDCLM model are computed by mixing the component distanced word probabilities for classes and the interpolated class information for histories. Similar to the DCLM model, the parameters of the IDCLM model are computed by using the variational Bayesian-EM procedure.

The rest of this paper is organized as follows. Section 2 is used for reviewing the DCLM model. The proposed IDCLM model is described in section 3. The comparison of the IDCLM and the DCLM models is described in section 4. The experimental details are described in section 5. Finally, the conclusions and future work are described in section 6.

2 DCLM

LDA is used to compute the document probability by using the topic structure at the document level, which is inconsistent with the language model for speech recognition where the n-gram regularities are characterized (Chien and Chueh, 2011). The DCLM was developed to model the n-gram events of the corpus for speech recognition. In the DCLM, the class structure is described by Dirichlet densities and estimated from n-gram events. The graphical model of the DCLM for a text corpus that comprises n-gram events $\{w_{i-n+1}^{i-1}, w_i\}$ is described in Figure 1. Here, H and N_h represent the number of history events w_{i-n+1}^{i-1} and the number of collected words that occur following the history w_{i-n+1}^{i-1}, respectively. The (n-1) history words w_{i-n+1}^{i-1} are represented by a (n-1)$V \times 1$ vector $\mathbf{h}$, consisting of n-1 block subvectors, with the entries of the seen words assigned to ones and those of unseen words assigned to zeros (Chien and Chueh, 2011). Here, V represents the size of the vocabulary. The vector $\mathbf{h}$ is then projected into a C-dimensional continuous class space using a class-dependent linear discriminant function:

$$g_c(\mathbf{h}) = \mathbf{a}_c^T \mathbf{h} \tag{1}$$

where $\mathbf{a}_c^T$ is the c^{th} row vector of matrix $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_C]$ (Chien and Chueh, 2011). The function $g_c(\mathbf{h})$ describes the class posterior probability $p(c|\mathbf{h})$, which is used in predicting the class information for an unseen history (Chien and Chueh, 2011). The model can be described as:

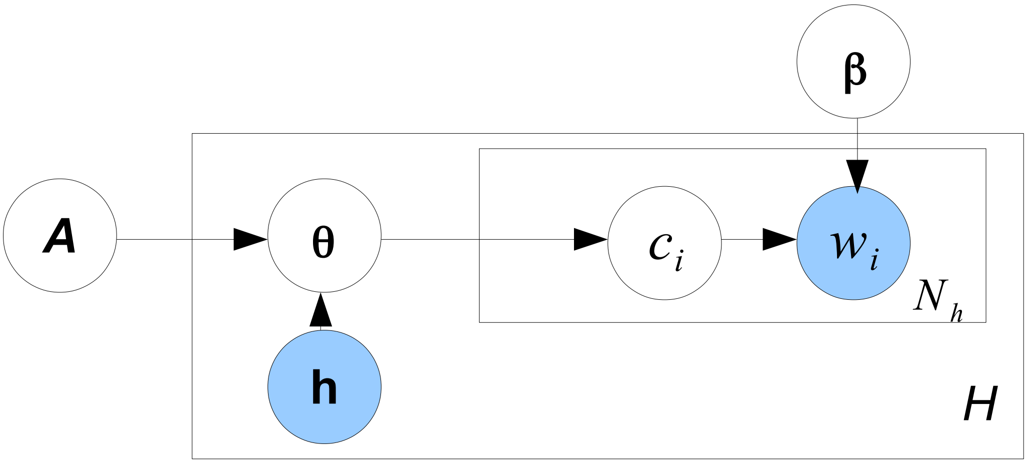

Figure 1: The graphical model of the DCLM. Shaded circles represent observed variables.

- For each history vector $\mathbf{h}$, the class information c is drawn from a history-dependent Dirichlet prior $\boldsymbol{\theta}$, which is related to a global projection matrix $\mathbf{A}$:

$$p(\boldsymbol{\theta}|\mathbf{h},\mathbf{A}) \propto \prod_{c=1}^{C} \boldsymbol{\theta}_c^{g_c(\mathbf{h})-1}, \tag{2}$$

- For each predicted word w_i of the n-gram events from a multinomial distribution with parameter $\boldsymbol{\beta}$, the associated class c_i is chosen by using a multinomial distribution with parameter $\boldsymbol{\theta}$. The joint probability of the variable $\boldsymbol{\theta}$, c_i, and w_i conditioned on $\mathbf{h}$ can be computed as:

$$p(\boldsymbol{\theta}, c_i, w_i|\mathbf{h},\mathbf{A},\boldsymbol{\beta}) = p(\boldsymbol{\theta}|\mathbf{h},\mathbf{A})p(c_i|\boldsymbol{\theta})p(w_i|c_i,\boldsymbol{\beta}) \tag{3}$$

- The conditional probability in the n-gram language model can thus be obtained as:

$$p(w_i|\mathbf{h},\mathbf{A},\boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\mathbf{h},\mathbf{A}) \sum_{c_i=1}^{C} p(c_i|\boldsymbol{\theta})p(w_i|c_i,\boldsymbol{\beta})d\boldsymbol{\theta}, \tag{4}$$

  where the integral is computed as:

$$p(c_i|\mathbf{h},\mathbf{A}) = \int p(\boldsymbol{\theta}|\mathbf{h},\mathbf{A})p(c_i|\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{g_{c_i}(\mathbf{h})}{\sum_{j=1}^{C} g_j(\mathbf{h})}. \tag{5}$$

  which is an expectation of a Dirichlet distribution of latent class c_i (Chien and Chueh, 2011).

Therefore, the probability of an n-gram event using the DCLM (Equation 4 and 5) can be written as (Chien and Chueh, 2011):

$$p(w_i|\mathbf{h},\mathbf{A},\boldsymbol{\beta}) = \sum_{c=1}^{C} p(w_i|c,\boldsymbol{\beta}) \frac{g_c(\mathbf{h})}{\sum_{j=1}^{C} g_j(\mathbf{h})} \tag{6}$$

The parameters ($\mathbf{A}$,$\boldsymbol{\beta}$) of the model are computed by using the variational bayesian EM (VB-EM) procedure (Chien and Chueh, 2011).

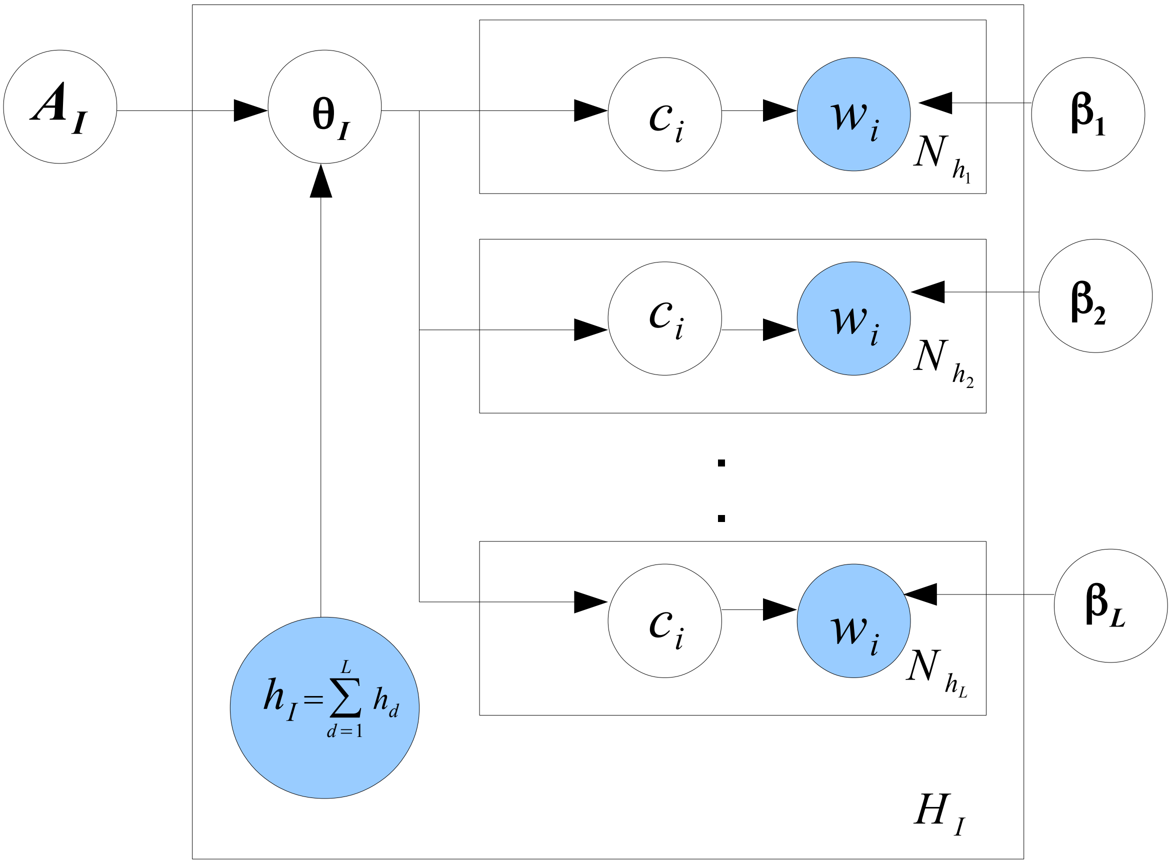

Figure 2: The graphical model of the IDCLM. Shaded circles represent observed variables.

3 Proposed IDCLM

The DCLM does not capture the long-range information from outside of the n-gram window (Chien and Chueh, 2011). To incorporate the long-range information into the DCLM, we propose an IDCLM where the class information is extracted from interpolated distance n-gram histories through a Dirichlet distribution in calculating the language model probability. In this model, we interpolate the distanced n-gram events into the original n-gram events of the DCLM. The graphical model of the IDCLM is described in Figure 2. In Figure 2, H_I contains the histories of all the distanced d n-grams, d represents the distance between words in the n-gram events, and L describes the maximum length of distance d. When d = 1, the n-grams are the default background n-grams. For example, the distanced tri-grams of the phrase "Interpolated Dirichlet Class Language Model for Speech Recognition" are described in Table 1 for the distance $d = 1, 2, 3$. Here, the (n-1)V dimensional discrete history vector $\mathbf{h}_I$ is projected into a C-dimensional continuous class space using a class-dependent linear discriminant function:

d	Trigrams
1	*Interpolated Dirichlet Class, Dirichlet Class Language, Class Language Model, Language Model for, Model for Speech, for Speech Recognition*
2	*Interpolated Class Model, Dirichlet Language for, Class Model Speech, Language for Recognition*
3	*Interpolated Language Speech, Dirichlet Model Recognition*

Table 1: *Distanced tri-grams for the phrase "Interpolated Dirichlet Class Language Model for Speech Recognition"*

$$g_c(\mathbf{h}_I) = \mathbf{a}_{c,I}^T \mathbf{h}_I \tag{7}$$

where $\mathbf{h}_I$ is the combined histories of all the distanced histories $\mathbf{h}_d$ and is defined as $\mathbf{h}_I = \sum_{d=1}^{L} \mathbf{h}_d$. Here, $\sum$ represents the logical OR operator. $\mathbf{a}_{c,I}^T$ is the c^{th} row vector of the matrix $\mathbf{A}_I$ and $g_c(\mathbf{h}_I)$ describes the class posterior probability $p(c|\mathbf{h}_I)$.

The n-gram probability of the IDCLM model is computed as:

$$\begin{aligned} p_I(w_i|\mathbf{h}_I, \mathbf{A}_I, \boldsymbol{\beta}_d) &= \sum_{c_i=1}^{C} \left\{ \Big[\sum_{d} \lambda_d p_d(w_i|c_i, \boldsymbol{\beta}_d) \Big] \times \int p(\boldsymbol{\theta}_I|\mathbf{h}_I, \mathbf{A}_I) p(c_i|\boldsymbol{\theta}_I) d\boldsymbol{\theta}_I \right\} \\ &= \sum_{c=1}^{C} \Big[\sum_{d} \lambda_d \beta_{d,ic} \Big] \frac{g_c(\mathbf{h}_I)}{\sum_{j=1}^{C} g_j(\mathbf{h}_I)} \end{aligned} \tag{8}$$

where λ_d are the weights for each component probability estimated on the held-out data using the EM algorithm (Bassiou and Kotropoulos, 2010; Dempster et al., 1977).

The parameters of the IDCLM model are computed using the variational Bayes EM (VB-EM) procedure by maximizing the marginal distribution of the training data that contains a set of n-gram events $D = \{w_{i-n+1}^{i-1}, w_i\}$:

$$\begin{aligned} \log p(D|\mathbf{A}_I, \boldsymbol{\beta}_d) &= \sum_{(w_i, \mathbf{h}_I) \in D} \log p_I(w_i|\mathbf{h}_I, \mathbf{A}_I, \boldsymbol{\beta}_d) \\ &= \sum_{\mathbf{h}_I} \log \left\{ \int p(\boldsymbol{\theta}_I|\mathbf{h}_I, \mathbf{A}_I) \times \left[\sum_{d} \prod_{j=1}^{N_{h_d}} \sum_{c_j=1}^{C} \lambda_d p_d(w_j|c_j, \boldsymbol{\beta}_d) p(c_j|\boldsymbol{\theta}_I) \right] d\boldsymbol{\theta}_I \right\} \end{aligned} \tag{9}$$

where D contains all the distanced n-gram events, N_{h_d} represents the number of collected words that occur following the history $\mathbf{h}_d$ in d-distanced n-grams. In Equation 9, the summation is over all possible histories in training samples D. However, directly optimizing the Equation 9 is intractable (Chien and Chueh, 2011). A variational IDCLM is introduced where the marginal likelihood is approximated by maximizing the lower bound of Equation 9. The VB-EM procedure is required since the parameter estimation involves the latent variables of $\{\boldsymbol{\theta}_I, \mathbf{c}_{h_d} = \{c_i\}_{i=1}^{N_{h_d}}\}$.

The lower bound $L(\mathbf{A}_I, \boldsymbol{\beta}_d; \hat{\boldsymbol{\gamma}}_I, \hat{\boldsymbol{\phi}}_d)$ is given by:

$$\begin{aligned} &\sum_{\mathbf{h}_I} \left\{ \log \Gamma\left(\sum_{c=1}^{C} g_c(\mathbf{h}_I) \right) - \sum_{c=1}^{C} \log \Gamma(g_c(\mathbf{h}_I)) + \sum_{c=1}^{C} (g_c(\mathbf{h}_I) - 1) \times \left(\Psi(\gamma_{h_I,c}) - \Psi\Big(\sum_{j=1}^{C} \gamma_{h_I,j} \Big) \right) \right\} \\ &\qquad + \sum_{d} \sum_{\mathbf{h}_d} \sum_{i=1}^{N_{h_d}} \sum_{c=1}^{C} \lambda_d \phi_{h_d,ic} \left(\Psi(\gamma_{h_I,c}) - \Psi\Big(\sum_{j=1}^{C} \gamma_{h_I,j} \Big) \right) \\ &\qquad + \sum_{d} \sum_{\mathbf{h}_d} \sum_{i=1}^{N_{h_d}} \sum_{c=1}^{C} \sum_{v=1}^{V} \lambda_d \phi_{h_d,ic} \delta(w_v, w_i) \log \beta_{d,vc} - \sum_{\mathbf{h}_I} \left\{ \log \Gamma\Big(\sum_{c=1}^{C} \gamma_{h_I,c} \Big) - \sum_{c=1}^{C} \log \Gamma(\gamma_{h_I,c}) \right. \\ &\qquad \left. + \sum_{c=1}^{C} (\gamma_{h_I,c} - 1) \left(\Psi(\gamma_{h_I,c}) - \Psi\Big(\sum_{j=1}^{C} \gamma_{h_I,j} \Big) \right) \right\} - \sum_{d} \sum_{\mathbf{h}_d} \sum_{i=1}^{N_{h_d}} \sum_{c=1}^{C} \lambda_d \phi_{h_d,ic} \log \phi_{h_d,ic} \end{aligned}$$

where $\Psi(.)$ is the derivative of the log gamma function, and is known as a digamma function (Chien and Chueh, 2011). The history-dependent variational parameters $\{\hat{\boldsymbol{\gamma}}_{h_I} = \hat{\gamma}_{h_I,c}, \hat{\boldsymbol{\phi}}_{h_d} = \hat{\phi}_{h_d,vc}\}$, corresponding to the latent variables $\boldsymbol{\theta}_I, \mathbf{c}_{h,d}$, are then estimated in the VB-E step by setting the differentials $(\partial L(\boldsymbol{\gamma}))/(\partial \gamma_{h_I,c})$ and $(\partial L(\boldsymbol{\phi}))/(\partial \phi_{h_d,ic})$ to zero respectively (Chien and Chueh, 2011):

$$\hat{\gamma}_{h_I,c} = g_c(\mathbf{h}_I) + \sum_{d} \sum_{i=1}^{N_{h_d}} \lambda_d \phi_{h_d,ic} \tag{10}$$

$$\hat{\phi}_{h_d,ic} = \frac{\beta_{d,ic}\exp\left[\Psi(\gamma_{h_I,c}) - \Psi(\sum_{j=1}^{C}\gamma_{h_I,j})\right]}{\sum_{l=1}^{C}\beta_{d,il}\exp\left[\Psi(\gamma_{h_I,l}) - \Psi(\sum_{j=1}^{C}\gamma_{h_I,j})\right]} \tag{11}$$

In computing $\hat{\phi}_{h_d,ic}$ the corresponding $\gamma_{h_d,c}$ is used in Equation 11. With the updated $\hat{\boldsymbol{\gamma}}_{h_I}$, $\hat{\boldsymbol{\phi}}_{h_d}$ in the VB-E step, the IDCLM parameters $\{\mathbf{A}_I,\boldsymbol{\beta}_d\}$ are estimated in the VB-M step as (Chien and Chueh, 2011):

$$\hat{\beta}_{d,vc} = \frac{\sum_{\mathbf{h}_d}\sum_{i=1}^{N_{h_d}}\lambda_d\hat{\phi}_{h_d,ic}\delta(w_v,w_i)}{\sum_{m=1}^{V}\sum_{\mathbf{h}_d}\sum_{i=1}^{N_{h_d}}\lambda_d\hat{\phi}_{h_d,ic}\delta(w_m,w_i)} \tag{12}$$

where $\sum_{v=1}^{V}\beta_{d,vc}$=1 and $\delta(w_v,w_i)$ is the Kronecker delta function that equals one when vocabulary word w_v is identical to the predicted word w_i and equals zero otherwise. The gradient ascent algorithm is used to calculate the parameters $\hat{\mathbf{A}}_I = [\hat{\mathbf{a}}_{1,I},\cdots,\hat{\mathbf{a}}_{C,I}]$ by updating the gradient $\bigtriangledown_{\mathbf{a}_{c,I}}$as (Chien and Chueh, 2011):

$$\bigtriangledown_{\mathbf{a}_{c,I}} \leftarrow \bigtriangledown_{\mathbf{a}_{c,I}} + \sum_{\mathbf{h}_I}\left[\Psi\Big(\sum_{j=1}^{C}g_j(\mathbf{h}_I)\Big) - \Psi(g_c(\mathbf{h}_I)) + \Psi(\hat{\gamma}_{h_I,c}) - \Psi\Big(\sum_{j=1}^{C}\hat{\gamma}_{h_I,j}\Big)\right].\mathbf{h}_I \tag{13}$$

The n-gram probabilities $p_t(w_i,\mathbf{h}_t,\mathbf{A}_I,\boldsymbol{\beta}_d)$ of the test document t are then computed using Equation 8. To capture the local lexical regularities, the model $p_t(w_i|\mathbf{h}_t,\mathbf{A}_I,\boldsymbol{\beta}_d)$ is then interpolated with the background trigram model as:

$$p_{Interpolated}(w_i|\mathbf{h}) = \mu p_{Background}(w_i|\mathbf{h}) + (1-\mu)p_t(w_i|\mathbf{h}_t,\mathbf{A}_I,\boldsymbol{\beta}_d) \tag{14}$$

4 Comparison of DCLM and IDCLM Models

In the DCLM model, the class information for the $(n-1)$ history words is obtained by using the n-gram counts in the corpus. The current word is predicted from the history-dependent Dirichlet parameter, which is controlled by a matrix $\mathbf{A}$ and corpus-based histories $\mathbf{h}$ (Chien and Chueh, 2011). In contrast, the IDCLM model captures long-range information by incorporating distanced n-grams. Here, the class information is exploited for the interpolated $(n-1)$ history words $\mathbf{h_I}$ that are obtained from all the distanced n-gram events. Both the DCLM and IDCLM exploit the word distribution given the history words. They perform the history clustering of the corpus. For the DCLM model, the number of parameters $\{\mathbf{A},\boldsymbol{\beta}\}$ increases linearly with the number of history words and is given by $(n-1)CV + CV$. For the IDCLM model, the number of parameters $\{\mathbf{A}_I,\boldsymbol{\beta}_d\}$ increases linearly with the number of history words and distance d and is given by $((n-1)CV + CVd)$. The time complexity of DCLM and IDCLM are $O(HVC)$ and $O(H_IVCd)$ with H corpus-based histories, H_I corpus-based interpolated histories, V vocabulary words, d distances and C classes.

5 Experiments

5.1 Data and experimental setup

The LM approaches are evaluated using the Wall Street Journal (WSJ) corpus (Paul and Baker, 1992). The SRILM toolkit (Stolcke, 2002) and the HTK toolkit (Young et al., 2013) are used for generating the LMs and computing the WER respectively. The '87-89 WSJ corpus is used to train language models. The background trigrams are trained using the back-off version of the Witten-Bell smoothing; the 5K non-verbalized punctuation closed vocabulary. We train the trigram IDCLM model using $L = 2$ and $L = 3$. Ten EM iterations in the VB-EM procedure were used. The initial values of the entries in the matrix $\boldsymbol{\beta}$,$\boldsymbol{\beta}_d$ were set to be 1/V and those in $\mathbf{A}$, $\mathbf{A}_I$ were randomly set in the range [0,1]. To update the variational parameters in the VB-E step, one iteration was used. The VB-M step was executed to update the parameters $\mathbf{A}$, $\mathbf{A}_I$ by three iterations (Chien and Chueh, 2011). To capture the local lexical regularity, trigrams of various methods are interpolated with the background trigrams. The acoustic model from (Vertanen, 2013) is used in our experiments. The acoustic model is trained by using all WSJ and TIMIT (Garofolo et al., 1993) training data, the 40-phone set of the CMU dictionary (-, 2013), approximately 10000 tied-states, 32 Gaussians per state and 64 Gaussians per silence state. The acoustic waveforms are parameterized into a 39-dimensional feature vector consisting of 12 cepstral coefficients plus the 0^{th} cepstral, delta and delta delta coefficients, normalized using cepstral mean subtraction $(MFCC_{0-D-A-Z})$. We evaluated the cross-word models. The values of the word insertion penalty, beam width, and the language model scale factor are -4.0, 350.0, and 15.0 respectively (Vertanen, 2013). The interpolation weights λ_d and μ are computed by optimizing on the held-out data according to the metric of perplexity. The experiments are evaluated on the evaluation test, which is a total of 330 test utterances from the November 1992 ARPA CSR benchmark test data for vocabularies of 5K words (Paul and Baker, 1992; Woodland et al., 1994).

5.2 Experimental Results

Due to the higher memory and training time requirements for the IDCLM model, we trained the DCLM and IDCLM models for class sizes of 10 and 20. The perplexity and WER results are described in Table 2 and Figure 3 respectively.

Language Model	10 Classes	20 Classes
Background (B)	109.41	109.41
B+Class	106.65	106.97
B+DCLM	100.20	100.45
B+IDCLM (L=2)	98.01	97.94
B+IDCLM (L=3)	95.63	95.43

Table 2: Perplexity results of the models

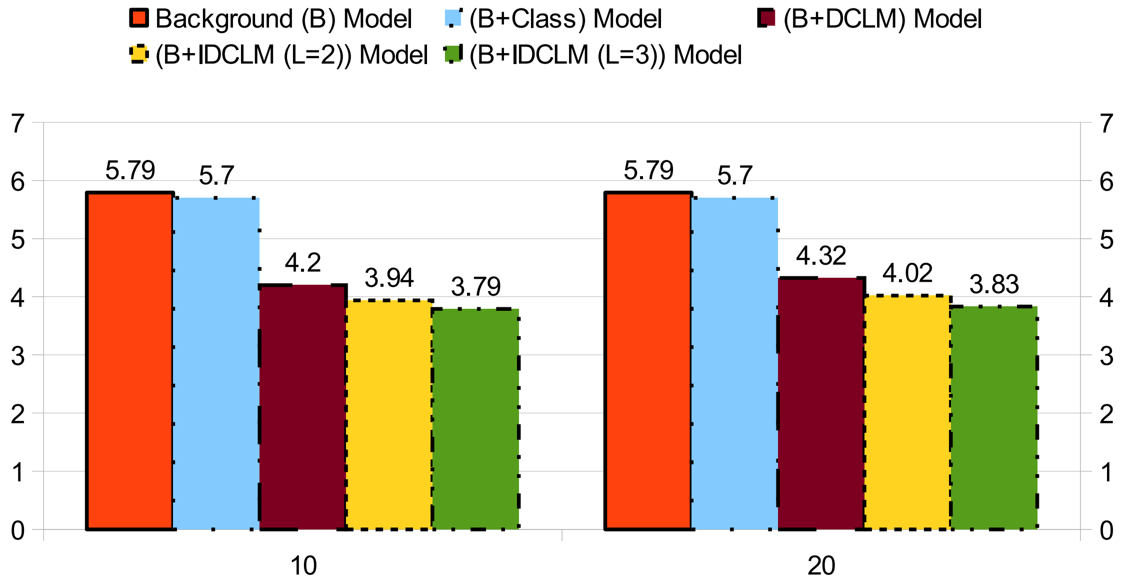

Figure 3: WER results for different class sizes

From Table 2, we can note the proposed IDCLM model outperforms the other models for all class sizes. The performance of IDCLM improves with more distances ($L = 3$).

We evaluated the WER experiments using lattice rescoring. In the first pass decoding, we used the background trigram for lattice generation. In the second pass, we applied the interpolated model for lattice rescoring. The WER results are described in Figure 3. From Figure 3, we can note that the proposed IDCLM ($L = 3$) model yields a WER reduction of about 34.54% (5.79% to 3.79%), 33.5% (5.7% to 3.79%), and 9.76% (4.2% to 3.79%) for 10 classes and about 33.85% (5.79% to 3.83%), 32.8%

(5.7% to 3.83%), and 11.34% (4.32% to 3.83%) over the background trigram, class trigram (Brown et al., 1992), and the DCLM (Chien and Chueh, 2011) approaches respectively. The significance improvement in WER is done by using a match-pair-test where the misrecognized words in each test utterance are counted. The p-values are described in Table 3. From Table 3, we can note that the IDCLM ($L = 2$) is statistically significant to the class-based LM (Brown et al., 1992) and DCLM (Chien and Chueh, 2011) at a significance level of 0.01 and 0.05 respectively. However, the IDCLM ($L = 3$) model is statistically significant to the above models at a significance level of 0.01. We have also seen that the cache DCLM model also gives the same results as DCLM (Chien and Chueh, 2011) for smaller number of classes (Chien and Chueh, 2011).

Language Model	10 Classes	20 Classes
B+Class & B+IDCLM (L=2)	3.8E-10	4.3E-10
B+Class & B+IDCLM (L=3)	4.7E-12	4.7E-12
B+DCLM & B+IDCLM (L=2)	0.04	0.01
B+DCLM & B+IDCLM (L=3)	0.004	0.006

Table 3: p-values obtained from the match-pair test on the WER results

6 Conclusions and Future Work

In this paper, we proposed an integration of distanced n-grams into the original DCLM model (Chien and Chueh, 2011). The DCLM model (Chien and Chueh, 2011) extracted the class information from the (n-1) history words through a Dirichlet distribution in calculating the n-gram probabilities. However, it does not capture the long-range semantic information from outside of the n-gram events. The proposed IDCLM overcomes the shortcomings of DCLM by incorporating the interpolated long-distance n-grams that capture the long-term word dependencies. Using the IDCLM, the class information for the histories is trained using the interpolated distanced n-grams. The IDCLM yields better results with including more distances ($L = 3$). The model probabilities are computed by weighting the component word probabilities for classes and the interpolated class information for histories. A variational Bayesian EM (VB-EM) procedure is presented to estimate the model parameters.

For future work, we will evaluate the proposed approach with neural network-based language models and exponential class-based language models. Furthermore, we will find out a way to perform the experiments for higher numbers of classes.

References

-. 2013. The Carnegie Mellon University (CMU) Pronounciation Dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B 39(1):1 – 38.

Andreas Stolcke. 2002. SRILM-an Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*, pages 901–904.

Chuang-H. Chueh and Jen-T. Chien. 2010. Topic Cache Language Model for Speech Recognition. In *Proc. of ICASSP*, pages 5194–5197.

Daniel Gildea and Thomas Hofmann. 1999. Topic-based Language Models using EM. In *Proceedings of EUROSPEECH*, pages 2167–2170.

David Mrva and Philip C. Woodland. 2004. A PLSA-based Language Model for Conversational Telephone Speech. In *Proc. of ICSLP*, pages 2257–2260.

David M. Blei, Andrew Y. Ng., and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dougls B. Paul and Janet M. Baker. 1992. The Design for the Wall Street Journal-based CSR Corpus. In *Proc. of ICSLP*, pages 899–902.

Hirofumi Yamamoto, Shuntaro Isogai, and Yoshinori Sagisaka. 2003. Multi-class Composite n-gram Language Model. *Speech Communication*, 41:369 – 379.

Holger Schwenk. 2007. Continuous Space Language Models. *Computer Speech and Language*, 21:492 – 518.

Jen-T. Chien and Chuang-H. Chueh. 2008. Latent Dirichlet Language Model for Speech Recognition. In *Proc. of IEEE SLT Workshop*, pages 201–204.

Jen-T. Chien and Chuang-H. Chueh. 2011. Dirichlet Class Language Models for Speech Recognition. *IEEE Trans. on Audio, Speech and Language Processing*, 19(3):482 – 495.

Jen-T. Chien. 2006. Association Pattern Language Modeling. *IEEE Trans. on Audio, Speech and Language Processing*, 14(5):1719 – 1728.

Jerome R. Bellegarda. 2000. Exploiting Latent Semantic Information in Statistical Language modeling. *IEEE Transactions on Speech and Audio Processing*, 88 (8):1279–1296.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. TIMIT Acoustic-phonetic Continuous Speech Corpus. *Linguistic Data Consortium.*

Keith Vertanen. 2013. HTK Wall Street Journal Training Recipe. `http://www.keithv.com/software/htk/us/`.

Md. A. Haidar and Douglas O'Shaughnessy. 2011. Unsupervised Language Model Adaptation using N-gram Weighting. In *Proceedings of CCECE*, pages 857–860.

Md. A. Haidar and Douglas O'Shaughnessy. 2012a. LDA-based LM Adaptation using Latent Semantic Marginals and Minimum Discrimination Information. In *Proceedings of EUSIPCO*, pages 2040–2044.

Md. A. Haidar and Douglas O'Shaughnessy. 2012b. Topic N-gram Count Language Model for Speech Recognition. In *Proceedings of IEEE Spoken Language Technology (SLT) Workshop*, pages 165–169.

Md. A. Haidar and Douglas O'Shaughnessy. 2013a. Fitting Long-range Information using Interpolated Distanced n-grams and Cache Models into a Latent Dirichlet Language Model for Speech Recognition. In *Proc. of INTERSPEECH*, pages 2678–2682.

Md. A. Haidar and Douglas O'Shaughnessy. 2013b. PLSA Enhance with a Long-distance Bigram Language Model for Speech Recognition. In *Proc. of EUSIPCO*.

Nikoletta Bassiou and Constantine Kotropoulos. 2010. Word Clustering PLSA Enhanced with Long Distance Bigrams. In *Proc. of International Conferance on Pattern Recognition*, pages 4226–4229.

P.C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. 1994. Large Vocabulary Continuous Speech Recognition using HTK. In *Proceedings of ICASSP*, pages 125–128.

Peng Xu and Frederick Jelinek. 2007. Random Forests and the Data Sparseness Problem in Language Modeling. *Computer Speech and Language*, 21 (1):105 – 152.

Peter F. Brown, Vincent Della Pietra, Peter De Souza, Jenifer Lai, and Robert L. Mercer. 1992. Classbased n-gram Models of Natural Language. *Computational Linguist.*, 18 (4):467 – 479.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for m-gram Language Modeling. In *Proc. IEEE Int Conf. Acoust., Speech, Signal Process.*, pages 181–184.

Roland Kuhn and Renato D. Mori. 1990. A Cache-based Natural Language Model for Speech Recognition. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 12 (6):570–583.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391 – 407.

Shuanghu Bai, Haizhou Li, Zhiwei Lin, and Baosheng Yuan. 1998. Building Class-based Language Models with Contextual Statistics. In *Proc. IEEE Int Conf. Acoust., Speech, Signal Process*, pages 173–176.

Slava M. Katz. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. *EEE Trans. Acoust., Speech, Signal Process.*, 35(3):400 – 401.

Stanley Chen, 2008. *Performance Prediction for Exponential Language Models*. Tech. Rep. RC 24671, IBM Research, Tech. Rep.

Steve Young, Phil Woodland, Gunnar Evermann, and Mark Gales. 2013. The HTK Toolkit 3.4.1. `http://htk.eng.cam.ac.uk/`.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 289–296, San Francisco, CA. Morgan Kaufmann.

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan H. Cernocky, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. In *Proc. of INTERSPEECH*, pages 1045–1048.

Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan H. Cernocky, and Sanjeev Khudanpur. 2011. Extensions Recurrent Neural Network Language Model. In *Proc. of ICASSP*, pages 5528–5531.

Welly Naptali, Masatoshi Tsuchiya, and Seiichi Nakagawa. 2012. Topic Dependent Class-based n-gram Language Model. *IEEE Trans. on Audio, Speech and Language Processing*, 20:1513 – 1525.

Yik-Cheung Tam and Tanja Schultz. 2005. Dynamic Language Model Adaptation using Variational Bayes Inference. In *Proceedings of INTERSPEECH*, pages 5–8.

Yik-Cheung Tam and Tanja Schultz. 2006. Unsupervised Language Model Adaptation using Latent Semantic Marginals. In *Proceedings of INTERSPEECH*, pages 2206–2209.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137 – 1155.