Fitting Long-range Information Using Interpolated Distanced N-grams and Cache Models into a Latent Dirichlet Language Model for Speech Recognition

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

Abstract

We propose a language modeling (LM) approach using interpolated distanced n-grams into a latent Dirichlet language model (LDLM) [1] for speech recognition. The LDLM relaxes the bag-of-words assumption and document topic extraction of latent Dirichlet allocation (LDA). It uses default background n-grams where topic information is extracted from the (n-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 events that can improve the language modeling performance. In this paper, we present an interpolated LDLM (ILDLM) by using different distanced n-grams. Here, the topic information is exploited from (n-1) history words through the Dirichlet distribution using interpolated distanced n-grams. The n-gram probabilities of the model are computed by using the distanced word probabilities for the topics and the interpolated topic information for the histories. In addition, we incorporate a cache-based LM, which models the re-occurring words, through unigram scaling to adapt the LDLM and ILDLM models that model the topical words. We have seen that our approaches give significant reductions in perplexity and word error rate (WER) over the probabilistic latent semantic analysis (PLSA) and LDLM approaches using the Wall Street Journal (WSJ) corpus.

Index Terms: language model, speech recognition, long-distance n-grams, topic models

1. Introduction

Statistical n-gram LMs have been successfully used for speech recognition and many other applications. The n-gram models suffer from the insufficiencies of the long-range information which limit a system’s performance. To capture the long-range information, one of the earliest attempts was a cache-based LM that took advantage 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 [2]. Recently, latent topic analysis has been used broadly to compensate for the weaknesses of n-gram models. Several techniques such as Latent Semantic Analysis (LSA) [3], PLSA [4], and LDA [5] have been studied to extract the latent semantic information from a training corpus. All these methods are based on a bag-of-words assumption. In LSA, semantic information can be obtained from a word-document co-occurrence matrix. In PLSA and LDA, semantic properties of words and documents can be shown in probabilistic topics. The PLSA latent topic parameters are trained by maximizing the likelihood of the training data using an expectation maximization (EM) procedure and have been successfully used for speech recognition [4, 6]. It is prone to the overfitting problem for a large number of documents, as each document has its own mixture weights. So, PLSA cannot be used to model an unseen document. On the other hand, LDA can be used to model an unseen document as it imposes a Dirichlet distribution over topic mixture weights corresponding to the documents in the corpus. The LDA model has been used successfully in recent research work for LM adaptation [7, 8, 9, 10, 11]. 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 [1], 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 [12].

In [13], a PLSA technique enhanced with long-distance bigrams was used to incorporate the long-term word dependencies in determining word clusters. This motivates us to use the long-distance n-grams using interpolation to induce the long-term word dependencies into the LDLM model. In this paper, we capture the long-range information into the LDLM using the interpolated distanced n-grams and cache based models. The n-gram probabilities of the proposed ILDLM model are computed by mixing the component distanced word probabilities for topics and the interpolated topic information for histories. Furthermore, we incorporate a cache-based LM into the LDLM and ILDLM models as the cache-based LM models different parts of the language than the topic models.

The rest of this paper is organized as follows. Section 2 is used to review the LDA and LDLM. The proposed ILDLM model is described in section 3. In section 4, the unigram scaling of the cache-based model to the topic models is explained. Section 5 describes the experiments. Finally, the conclusion and future work are explained in section 6.

2. LDA and LDLM

2.1. LDA

LDA is a three-level hierarchical Bayesian model, where each item of a collection of discrete data is modeled as a finite mixture over an underlying set of topics. Each topic is then modeled as an infinite mixture over an underlying set of topic probabilities [5]. The model can be described as follows:

- Each document \( D = w_1, \ldots, w_N \) is generated as a mixture of unigram models, where the topic mixture weight \( \theta \) is drawn from a prior Dirichlet distribution:

\[
  f(\theta; \alpha) \propto \prod_{z=1}^{Z} \theta_{z}^{\alpha_{z} - 1}.
\]

- For each word in document \( D \):

\[
  f(w_i; \theta) \propto \sum_{z=1}^{Z} \theta_{z}^{w_{i,z}}.
\]

The rest of this paper is organized as follows. Section 2 is used to review the LDA and LDLM. The proposed ILDLM model is described in section 3. In section 4, the unigram scaling of the cache-based model to the topic models is explained. Section 5 describes the experiments. Finally, the conclusion and future work are explained in section 6.
2.2. LDLM

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 [1]. The LDLM were developed to model the n-gram events for speech recognition. The graphical model for LDLM is described in Figure 1. Here, $H$ and $V$ represent the number of histories and the size of vocabulary, respectively [1]. In this model, the topic mixture vector $\theta$ is generated by the history-dependent Dirichlet parameter. A parameter matrix $A$ is merged in the Dirichlet model to consider regularities are characterized [1]. The LDLM were developed to capture the long-range information from outside of the n-gram events [1]. To incorporate the long-range information into the LDLM, we propose an ILDLM where the topic 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 $d$-th n-gram event into the original n-gram events of the LDLM. The n-gram probability using the ILDLM can be defined as [13]:

$$P_{ILDLM}(w_i|h) = \frac{\sum_{z=1}^{Z} \lambda_d P_d(w_i|z, \beta_d)}{\sum_{j=1}^{Z} \frac{a^T_j h}{\sum_{j=1}^{Z} a^T_j h}}.$$  

(8)

where $\lambda_d$ are the weights for each component probability estimated on the held-out data using the EM algorithm. $d$ represents the distance between words in the n-gram events. $d = 1$ describes the default n-grams. For example, the distanced n-grams of the phrase “Speech in Life Sciences and Human Societies” are described in Table 1 for the distance $d = 1, 2$. 

Table 1: Distanced n-grams for the phrase “Speech in Life Sciences and Human Societies”

<table>
<thead>
<tr>
<th>Distance</th>
<th>Bigrams</th>
<th>Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d=1$</td>
<td>Speech in, Life Sciences, Sciences and, and Human, Human Societies</td>
<td>Speech in, Life Sciences, Life Sciences and, Sciences and Human, and Human Societies</td>
</tr>
<tr>
<td>$d=2$</td>
<td>Speech Life and, Sciences, Life and, Sciences, Human, and Societies</td>
<td>Speech Life and, in Sciences Human, Life and Societies</td>
</tr>
</tbody>
</table>

The parameters of the ILDLM model are computed using the EM procedure by maximizing the marginal distribution of the training data:

$$\sum_{h, w_i} n_d(h, w_i) \log P_{ILDLM}(w_i|h),$$  

(9)

where $n_d(h, w_i)$ are the distanced n-grams. In the E-step, the auxiliary function of the new estimates $A', \beta_d'$ given current estimates $A, \beta_d$ is calculated by taking the expectation of the
marginal likelihood function of Equation 9 over the hidden variable $z$ [1]:

$$Q(A', \beta'_d|A, \beta_d) = \sum_{h, v', d} n_d(h, w_{i'}) \lambda_d E_z[\log P_d(w_{i'}, z|h, A', \beta'_d)|A, \beta_d]$$

$$= \sum_{h, v', d} n_d(h, w_{i'}) \sum_{z=1}^{Z} \lambda_d P_d(z|h, w_{i'}, A, \beta_d) \log P_d(w_{i'}, z|h, A', \beta'_d)$$

$$= \sum_{h, v', d} n_d(h, w_{i'}) \sum_{z=1}^{Z} \lambda_d P_d(z|h, w_{i'}, A, \beta_d) \times$$

$$\log[P_d(w_{i'}|z, \beta'_d)] = \frac{\hat{a}^T_h}{\sum_{j=1}^{Z} a^T_j h},$$

where $P_d(z|h, w_{i'}, A, \beta_d)$ are the posterior probabilities of the latent variables which are calculated based on the current estimates as:

$$P_d(z|h, w_{i'}, A, \beta_d) = \frac{P_d(w_{i'}|z, \beta'_d) a^T_j h}{\sum_{j=1}^{Z} P_d(w_{i'}|j, \beta'_d) a^T_j h}$$

In the M-step, the new estimates $\beta'_d$ and $A'$ are computed. The parameter $\beta'_z, w_{i'}, d$ is updated as:

$$\hat{\beta}'_{z, w_{i'}, d} = \frac{\hat{a}^T_h}{\sum_{i'} \sum_{z=1}^{Z} n_d(h, w_{i'}) P_d(z|h, w_{i'}, A, \beta_d)}$$

To compute the parameter $A'$, the gradient ascent algorithm is used for maximization [14]. The gradient of the auxiliary function $\nabla_{\alpha'} Q(A', \beta'_d|A, \beta_d)$ is given by

$$\nabla_{\alpha'} Q(A', \beta'_d|A, \beta_d) = \sum_{h, v', d} n_d(h, w_{i'}) \lambda_d P_d(z|h, w_{i'}, A, \beta_d) \left[ \frac{1}{a^T_h} - \frac{1}{\sum_{j=1}^{Z} a^T_j h} \right] h$$

Therefore, the new parameter $a^T_h$ at the $(t + 1)$ iteration is updated by:

$$a^T_{h(t+1)} = a^T_{h(t)} + \eta \nabla_{a'} Q(A', \beta'_d|A, \beta_d).$$

where $\eta$ is the learning parameter. The model parameters are then estimated with several EM iterations.

4. Incorporating Cache Models into LDLM and ILDLM Models

A Cache-based language model was used to increase the probability of words appearing earlier in a document that are likely to occur in the same document. The unigram cache model for a given history $h_c = w_{1-M}, \ldots, w_i$, where $M$ is the cache size, is defined as:

$$P_{CACHE}(w_i) = \frac{C(w_i, h_c)}{C(h_c)},$$

where $C(w_i, h_c)$ is the number of occurrences of the word $w_i$ within $h_c$, and $C(h_c)$ is the number of words within $h_c$ that belong to the vocabulary $V$ [15].

The LDLM/ILDLM capture topical words. To capture the local lexical regularities, the models are interpolated with a background $n$-gram model as:

$$P_L(w_i|h) = \gamma P_{Background}(w_i)+ (1-\gamma) P_{LDLM/iLDLM}(w_i|h)$$

The cache-based LM models re-occurring words that are different from the background model (i.e., models short-range information), LDLM and ILDLM models (i.e., models topical words). Therefore, the cache model can be used to adapt the model $P_L(w_i|h)$ using unigram scaling as [16, 17]:

$$P_{Adapt}(w_i|h) = \frac{P_L(w_i|h) \delta(w_i)}{Z(h)}$$

with

$$Z(h) = \sum_{w_i} \delta(w_i).P_L(w_i|h)$$

where $Z(h)$ is a normalization term, which guarantees that the total probability sums to unity and $\delta(w_i)$ is a scaling factor, which is usually approximated as:

$$\delta(w_i) \approx \left( \frac{\rho P_{\text{CACHE}}(w_i) + (1 - \rho) P_{\text{Background}}(w_i)}{P_{\text{Background}}(w_i)} \right)^\mu,$$

where $\mu$ is a tuning factor between 0 and 1. In our experiments we used the value of $\mu$ as 1. We used the same procedure as [16] to compute the normalization term. To do this, an additional constraint is employed where the total probability of the observed transitions is unchanged:

$$\sum_{w_i, \text{observed}(h, w_i)} P_{Adapt}(w_i|h) = \sum_{w_i, \text{observed}(h, w_i)} P_L(w_i|h).$$

The model $P_L(w_i|h)$ has standard back-off structure and the above constraint, so the model $P_{Adapt}(w_i|h)$ has the following recursive formula:

$$P_{Adapt}(w_i|h) = \begin{cases} \frac{\delta(w_i)}{Z(h)} P_L(w_i|h) \text{if } (h, w_i) \text{ exists(18)} \\ b(h).P_{Adapt}(w_i|h) \text{ otherwise} \end{cases}$$

where

$$Z(h) = \sum_{w_i, \text{observed}(h, w_i)} \delta(w_i).P_L(w_i|h)$$

and

$$b(h) = 1 - \sum_{w_i, \text{observed}(h, w_i)} P_L(w_i|h)$$

where $b(h)$ is the back-off weight of the context $h$ to ensure that $P_{Adapt}(w_i|h)$ sums to unity. $h$ is the reduced word history of $h$. The term $Z(h)$ is used to do normalization similar to Equation 16 except the summation is considered only on the observed alternative words with the equal word history $h$ in the LM [8]. We describe the adaptation using the unigram scaling of cache models as (Background+LDLM/ILDLM)*CACHE.

5. Experiments

5.1. Data and experimental setup

The LM adaptation approaches are evaluated using the Wall Street Journal (WSJ) corpus [18]. The SRILM toolkit [19] and the HTK toolkit [20] are used for generating the LMs and computing the WER respectively. The ‘87-89 WSJ corpus is used to train language models. The models are trained using the WSJ 5K non-verbalized punctuation closed vocabulary. A trigram background model is trained using the modified Kneser-Ney smoothing incorporating the cutoffs 1 and 3 on the bi-gram and tri-gram counts respectively. To reduce the computational and memory requirements using MATLAB, we trained only the
bi-gram LDLM and ILDLM models. For ILDLM models, we considered bigrams for $d = 1, 2$. The learning parameter $\eta$ is set to 0.01. A fixed cache size of $M = 400$ is used for the cache-based LM. The acoustic model from [21] is used in our experiments. The acoustic model is trained by using all WSJ and TIMIT [22] training data, the 40 phone set of the CMU dictionary [23], 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 [21]. The development and the evaluation test sets are the si_dtl05.odd (248 sentences from 10 speakers) and the Nov'93 Hub 2.5K test data from the ARPA November 1993 WSJ evaluation (215 sentences from 10 speakers) [18, 24]. The interpolation weights $\lambda_n$, $\gamma$, and $\rho$ are computed using the compute-best-mix program from the SRILM toolkit. They are tuned on the development test set. The results are noted on the evaluation test set.

5.2. Experimental Results

We keep the unigram (Equations 7 and 8) probabilities for topics of LDLM and ILDLM, and $\lambda_n$ of component probabilities for ILDLM unchanged, and used them to compute the matrix $A$ for the test document’s histories [4]. The language models for LDLM and ILDLM are then computed using (Equations 7 and 8). The models are then interpolated with a back-off trigram background model to capture the local lexical regularities. Furthermore, a cache-based LM that models re-occurring words is integrated through unigram scaling (Equations 18 and 19) with the LDLM/ILDLM that models topical words. We also show the results for PLSA models using unigram scaling where the PLSA unigrams are used in place of cache unigrams in Equation 17 and denoted as Background+PLSA [4].

We tested the proposed approaches for various sizes of topics. The perplexity results of the experiments are described in Table 2. From Table 2, we can note that all the models outperform the background model and the performances are better with increasing topics. However, the proposed ILDLM model outperforms the PLSA and LDLM models in all forms (stand-alone, interpolated and unigram scaling) for all topic sizes. We evaluated the WER experiments using lattice rescoring. In the first pass, we used the back-off trigram background language model for lattice generation. In the second pass, we applied the LM adaptation approaches for lattice rescoring. The experimental results are explained in Figure 2. From Figure 2, we can note that the proposed ILDLM model yields significant WER reductions of about 20.28% (7.59% to 6.05%), 18.24% (7.40% to 6.05%), and 12.82% (6.94% to 6.05%) for 40 topics and about 22.00% (7.59% to 5.92%), 19.78% (7.38% to 5.92%), and 12.03% (6.73% to 5.92%) for 80 topics, over the background model, PLSA model [4], and the LDLM [1] approaches respectively. The integration of cache-based models improves the performance as it carries different information (captures the dynamics of word occurrences in a cache) than the LDLM and ILDLM approaches. The cache unigram scaling of the ILDLM approach gives 9.80% (6.63% to 5.98%) and 8.13% (6.39% to 5.87%) WER reductions over the cache unigram scaling of the LDLM approach for 40 and 80 topics respectively. We can note that the addition of cache models improves the performance of LDLM (6.94% to 6.63% for 40 topics and 6.73% to 6.39% for 80 topics) more than for ILDLM (6.05% to 5.98% for 40 topics and 5.92% to 5.87% for 80 topics). This might be due to fact that the ILDLM approach captures long-range information using the interpolated distanced bigrams. Therefore, it is proved that the proposed ILDLM approach includes long-range information into the LDLM model.

6. Conclusions and future work

In this paper, we proposed an integration of distanced $n$-grams into the original LDLM model [1]. The LDLM model extracted the topic information from the $(n-1)$ history words through a Dirichlet distribution in calculating the $n$-gram probabilities. However, it does not capture the long-term semantic information from outside of the $n$-gram events. The proposed ILDLM overcomes the shortcomings of LDLM by using the interpolated long-distance $n$-grams that capture the long-term word dependencies. Using the ILDLM, the topic information for the histories is trained using the interpolated distanced $n$-grams. The model probabilities are computed by weighting the component word probabilities for topics and the interpolated topic information for histories. We have seen that the proposed ILDLM approach yields significant perplexity and WER reductions over the LDLM approach using the WSJ corpus. Moreover, we incorporate a cache-based model into the topic models using unigram scaling for adaptation and have seen improved performances over the topic models. However, cache unigram scaling of the LDLM gives much better performance than the cache unigram scaling of the ILDLM. This proves that the proposed ILDLM approach captures long-range information of the language. For future work, we will apply the proposed approach for the Dirichlet class language model [25] using larger vocabulary and higher order $n$-grams.

Table 2: Perplexity results of the language models

<table>
<thead>
<tr>
<th>Language Model</th>
<th>40 Topics</th>
<th>80 Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>70.26</td>
<td>70.26</td>
</tr>
<tr>
<td>PLSA</td>
<td>517.77</td>
<td>514.78</td>
</tr>
<tr>
<td>LDLM</td>
<td>251.57</td>
<td>153.60</td>
</tr>
<tr>
<td>ILDLM</td>
<td>86.94</td>
<td>65.25</td>
</tr>
<tr>
<td>Background+PLSA</td>
<td>66.63</td>
<td>66.30</td>
</tr>
<tr>
<td>Background+LDLM</td>
<td>65.12</td>
<td>62.49</td>
</tr>
<tr>
<td>Background+ILDLM</td>
<td>53.61</td>
<td>52.69</td>
</tr>
<tr>
<td>(Background+LDLM)^CACHE</td>
<td>59.92</td>
<td>57.51</td>
</tr>
<tr>
<td>(Background+ILDLM)^CACHE</td>
<td>49.33</td>
<td>48.30</td>
</tr>
</tbody>
</table>
7. References


