Novel Weighting Scheme for Unsupervised Language Model Adaptation Using Latent Dirichlet Allocation

Md. Akmal Haidar and Douglas O’Shaughnessy
INRS-Énergie-Matériaux-Télécommunications, Montréal, Canada
haidar@emt.inrs.ca, dougo@emt.inrs.ca

Abstract

A new approach for computing weights of topic models in language model (LM) adaptation is introduced. We formed topic clusters by a hard-clustering method assigning one topic to one document based on the maximum number of words chosen from a topic for that document in Latent Dirichlet Allocation (LDA) analysis. The new weighting idea is that the unigram count of the topic generated by hard-clustering is used to compute the mixture weights instead of using an LDA latent topic word count used in the literature. Our approach shows significant perplexity and word error rate (WER) reduction against the existing approach.

Index Terms: latent dirichlet allocation, language model adaptation, mixture models, speech recognition.

1. Introduction

Language model (LM) adaptation plays an important role for many research areas like speech recognition, machine translation, and information retrieval etc. Adaptation is required when the styles, domains or topics of the test data are mismatched with the training data. It is also important as natural language is highly variable since the topic information is highly non-stationary. In general, an adaptive language model seeks to maintain an adequate representation of the domain under changing conditions involving potential variations in vocabulary, content, syntax and style [1]. The idea of an unsupervised LM adaptation approach is to extract the latent topics from the training set and then adapt the topic specific LM with proper mixture weights, finally interpolated with the generic n-gram LM.

Short range information can be captured through n-gram modeling. N-gram models use the local context information by modeling text as a Markovian Sequence. However, the training data is made out of a diverse collection of topics for which it is necessary to handle long-range information. In supervised LM adaptation, topic information of the training data is available; topic specific language models are then interpolated with the baseline language model [2, 3]. On the other hand, topic information is not available for unsupervised LM adaptation. There are various techniques to extract the latent semantic information from a training corpus such as Latent Semantic Analysis (LSA) [4], Probabilistic Latent Semantic Analysis (PLSA) [5], and LDA [6]. All the methods are based on a bag-of-words assumption, i.e., the word-order in a document can be ignored. 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. Here, the idea is that a document is formed as a mixture of topics and a topic is a probability distribution over words. PLSA is prone to the overfitting problem for a large number of documents, as each document that has its own mixture weights. So, it 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. However, the LDA model can be viewed as a mixture of unigram latent topic models.

The simple technique to extract a topic from an unlabeled corpus is to assign one topic label to a document [7]. In [8], this strategy is used with leveraging named entity information and LDA to form topics. Here topic-specific n-gram language models are created, weighted for adaptation and then the adapted model is interpolated with the baseline model, unlike all previous work where a unigram adapted topic model is interpolated with the baseline model. Weights for the topic language models are created using the LDA latent topic word count, which is appropriate for LDA unigram latent topic models. In this paper, we propose the idea that the weights of topic models are generated using the word count of the topics generated by a hard-clustering method instead of using the LDA latent topic word count and we have seen that our weighting approach gives better results in perplexity and WER reduction. It should be mentioned here that we used the same technique for creating topics and adapting topic models as in [8].

The rest of this paper is organized as follows. In section 2, related work on LDA and unsupervised language model adaptation is reviewed. Section 3 is used for describing the proposal and adaptation methodology. In section 4, experiments and results are explained. Finally the conclusion and future work are described in section 5.

2. Related work

Many methods have been investigated in unsupervised LM adaptation. An earlier attempt is a cache-based model, which was based on the concept that words that appear in a document are likely to occur again [9]. This idea is used in trigger-based LM adaptation using a maximum entropy approach [10]. Recently latent topic analysis has been introduced in language modeling. Topic clusters can be formed by assigning a single document to a topic and used in LM adaptation [7]. Bellegarda [4] extracted topics using Latent Semantic Analysis (LSA) and showed improvement in perplexity and WER reduction in LM adaptation. The probabilistic extension of LSA is PLSA, where topics are clustered in probabilistic space and have shown perplexity reduction in LM adaptation [5]. One of the most powerful probabilistic bag-of-words models is the LDA model. It overcomes the pitfalls of PLSA, which suffers from an over-fitting problem, and achieves better results in perplexity reduction [6].

LDA has been successfully used in recent research work in LM adaptation. Unigram topic models extracted by LDA models are dynamically adapted for unseen documents and interpolated with a background model [11]. In [12, 13], the MAP estimates of LDA parameters from variational lower bounds [6] are used for adjusting unigram probabilities. Style and topic adaptation are investigated using HMM-LDA [14]. LDA is also used as a clustering algorithm to cluster training
LDA is a popular probabilistic bag-of-words model [6]. It is a generative probabilistic model of text corpora, a collection of discrete data. LDA is a three level hierarchical Bayesian model, where each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is in turn modeled as an infinite mixture over an underlying set of topic probabilities. 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:

  \[
  f(\theta; \alpha) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k-1}
  \]  

- For each word in document \(d\):
  - Choose a topic \(k\) from the multinomial distribution \(\theta(d)\)
  - Choose a word \(w\) from the multinomial distribution \(\phi(w|k,\beta)\)

where \(\alpha=(\alpha_1,\ldots,\alpha_K)\) is used as the representation count for the \(K\) latent topics, \(\theta\) indicates the relative importance of topics for a document and \(\phi(w|k,\beta)\) represents the word probabilities conditioned on the topic with a Dirichlet prior and indicates the relative importance of particular words in a topic. \(\alpha\) and \(\beta\) are Dirichlet priors that control the smoothing of the topic distribution and topic-word distribution, respectively [17].

As a bag-of-word generative model, LDA assigns the following probability to a document \(d=w_1,\ldots,w_n\) as:

\[
p(d) = \int_{\theta} \prod_{i=1}^{n} \phi(w_i|k,\beta)\theta_k f(\theta;\alpha) d\theta
\]  

3.2. Topic clustering and language model generation

We have used the MATLAB topic modeling toolbox [19] to get the word-topic matrix, \(WP\), and the document-topic matrix, \(DP\), using LDA. Here, the words correspond to the words used in LDA analysis. In the \(WP\) matrix, an entry \(WP(j,k)\) represents the number of times word \(w_j\) has been assigned to topic \(z_i\) over the training set. In the \(DP\) matrix, an entry \(DP(i,k)\) contains the counts of words in document \(d_i\) that are from a topic \(z_i\) \((k=1,2,\ldots,K)\).

For training, topic clusters are formed by assigning a topic \(z_i^*\) to a document \(d_i\) as:

\[
z_i^* = \arg \max_{1 \leq k \leq K} DP(i, k)
\]  

i.e., a document is assigned to a topic from which it takes the maximum number of words. Therefore all the words of training documents are assigned to \(K\) topics. Then \(K\) topic N-gram LM’s are trained.

3.3. Language model adaptation

According to LDA, a document can be generated by a mixture of topics. So, for a test document \(d=w_1,\ldots,w_n\), we can create a dynamically adapted topic model by using a mixture of LMs from different topics as:

\[
P_{\text{LDA-adapt}}(w_i|h_k) = \sum_{i=1}^{K} \gamma_i p_{\theta}(w_i|h_k)
\]  

where \(p_{\theta}(w_i|h_k)\) is the \(i^{th}\) topic model and \(\gamma_i\) is the \(i^{th}\) mixture weight.

To find topic mixture weight \(\gamma_i\), we propose a scheme where the unigram count of the topics, generated by Equation (3) is used in place of the LDA topic word count matrix \(WP(j,k)\) used in [8]. This is because we used topics created by a hard-clustering strategy instead of using unigram latent topics extracted by LDA. Therefore,

\[
\gamma_k = \frac{\sum_{j=1}^{n} P(z_i|w_j)P(w_j|d)}{\sum_{j=1}^{n} P(z_i|w_j)P(w_j|d)}
\]  

\[
P(z_i|w_j) = \frac{TF(j,k)}{\sum_{p=1}^{K} TF(j,p)}
\]

\[
P(w_j|d) = \frac{freq(w_j)}{\sum_{q=1}^{n} freq(w_q)}
\]

where \(TF(j,k)\) represents the number of times word \(w_j\) is drawn from topic \(z_i\), which is created by Equation (3). \(freq(w_j)\) is the frequency of word \(w_j\) in document \(d\).

The adapted topic model is then interpolated with the generic LM as:

\[
P(w_i|h_k) = \lambda \ast P_{\text{glm}}(w_i|h_k) + (1-\lambda) \ast P_{\text{LDA-adapt}}(w_i|h_k)
\]

4. Experiments and results

4.1. Data and experimental setup

We evaluated the LM adaptation approach using the Brown Corpus and WSJ1 corpus transcription text data. For the Brown corpus, we have selected 100 documents randomly to form the test set and the remaining documents are used for training. For WSJ1, we used all the training transcription text data for training and development and evaluation test set 1 for testing. Here, we keep those sentences of the test set where all the words of the sentences are in the dictionary. We split the training transcription text data into 300 sentences per document and in total 261 documents are created. The summary of training and testing datasets is given in Table 1.

We used the SRILM toolkit [20] and HTK toolkit [21] for our experiments. We trained LMs and used the compute-best-mix program from the SRILM toolkit to find the optimal weight \(\lambda\) for interpolation with the background model. We used perplexity and WER to measure the performance of our
experiments. We used the baseline acoustic model from [22], where the model is trained by using all WSJ and TIMIT training data, the 40 phones set of the CMU dictionary, approximately 10000 tied-states, 32 gaussians per state and 64 gaussians per silence state. Here the acoustic waveforms are parameterized into a 39-dimensional feature vector consisting of 12 cepstral coefficients plus the 0th 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 beam width, word insertion penalty, and the language model scale factor are 350.0, -4.0, and 15.0 respectively.

Table 1: Summary of the Data Set

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of Words for Training</th>
<th>Number of Words for Testing</th>
<th>Size of Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown Corpus</td>
<td>926,946</td>
<td>234,112</td>
<td>49163</td>
</tr>
<tr>
<td>WSJ1 Transcription</td>
<td>1,317,793</td>
<td>Dev. Set:7235 Eval. Set:6708</td>
<td>20000</td>
</tr>
</tbody>
</table>

We performed experiments using the LDA latent topic word count weighting scheme [8] and using our proposed scheme. Latent topic analysis is done using LDA and topics are formed by Equation (3). This helps to assign all words of all documents into topics. The topics are then adapted for test data using Equations (4) and (5). The adapted topic model is interpolated with the background model using Equation (6).

4.2. Perplexity reduction

We carried out experiments for tri-grams and bi-grams on the Brown Corpus. As a single topic adapted model gives the worst performance in finer-grained topic analysis [8], we only show results with mixture topic models. We have used a maximum of 9 mixtures (limitation of the current SRILM toolkit) when the number of topics is greater than 9. We performed the LDA analysis on the training set of the Brown Corpus, which clusters the training set into topics. Then n-gram topic models are trained. The mixture weights are then found using Equation (5) and the adapted model is created using Equation (4). The LDA adapted model is then interpolated with the baseline model to obtain a better perplexity result. The results of the tri-gram and bi-gram models for optimal weight λ and a 50 topics training set are shown in Table 2. From Table 2, we can see that the interpolated LM created by both of these weighting techniques outperforms the background model with a proper interpolation weight. The perplexities of the test set using the baseline tri-gram and bi-gram model are 399.11 and 424.94. The best perplexity results achieved by the LDA latent topic word count weighting are 378.03 (λ=0.72) and 406.36 (λ=0.74) and by the proposed weighting scheme are 372.67 (λ=0.55) and 401.37 (λ=0.53) for tri-gram and bi-gram models respectively. From the table, we can observe that in every case the LDA latent topic word count weighting scheme requires bigger interpolation weight for the background model to achieve the best results. This is obvious since the topics are generated by hard clustering using Equation (3). We can also notice that, for tri-gram and bi-gram models with optimal weight λ, the interpolated model created by our proposed weighting scheme achieves perplexity reduction of about 6.62% and 5.54% against the background model and about 1.41% and 1.22% against the interpolated model formed by the LDA latent topic word count weighting scheme [8]. We can perform experiments for different numbers of topics. The results plotted in Figure 1 are obtained by using an optimal interpolation weight λ for the tri-gram model. From Figure 1, we can see that the perplexity is significantly reduced with an increasing number of topics. We can also observe that the proposed method requires smaller interpolation weights and outperforms the LDA latent topic word count weighting scheme in every case. When the topic number is 50, our proposed method achieves the lowest perplexity. After that, for example 75, the perplexity is reduced slightly. This might be due the limitation of the number of the topic mixtures used [8].

Table 2: Perplexity results of the N-gram model for optimal mixture weight λ for 50 topic clusters

<table>
<thead>
<tr>
<th>Language Model</th>
<th>N-gram</th>
<th>Optimal Mixture Weight λ</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Tri-gram</td>
<td>1.00</td>
<td>399.11</td>
</tr>
<tr>
<td></td>
<td>Bi-gram</td>
<td>1.00</td>
<td>424.94</td>
</tr>
<tr>
<td>Interpolated Model</td>
<td>Tri-gram</td>
<td>0.72</td>
<td>378.03</td>
</tr>
<tr>
<td>(LDA latent topic word count weighting)</td>
<td>Bi-gram</td>
<td>0.74</td>
<td>406.36</td>
</tr>
<tr>
<td>Interpolated Model</td>
<td>Tri-gram</td>
<td>0.55</td>
<td>372.67</td>
</tr>
<tr>
<td>(Proposed Scheme)</td>
<td>Bi-gram</td>
<td>0.53</td>
<td>401.37</td>
</tr>
</tbody>
</table>

Figure 1: Perplexity results of the test set for LDA latent topic word count and the proposed scheme using different numbers of topics and optimal mixture weights of the tri-gram model.

To evaluate WER using the HTK toolkit, we also trained the bi-gram model using WSJ1 training transcription text data. We employed LDA to create topic clusters. We formed 40 topics and the bi-gram topic models are created using the SRILM toolkit. Then the mixture weights are computed using Equation (5). The LDA adapted model is formed using Equation (4). Finally, the LDA adapted model is interpolated with the baseline bi-gram model, which is formed using training transcription text data. The results of the experiments are shown in Table 3. Again we note that both the weighting approaches outperform the baseline model and the LDA latent topic word count weighting scheme requires bigger
interpolation weight than the proposed scheme. The proposed approach achieves perplexity reductions of about 27.8% and 28% against the baseline model, about 1.1% and 1.9% against the LDA latent topic word count weighting scheme for the WSJ1 development test set 1 and evaluation test set 1.

Table 3: Perplexity results of the bi-gram model for Development and Evaluation Test sets using WSJ1 training transcription text.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Optimal Mixture Weight ( \lambda )</th>
<th>Perplexity (Development test set 1)</th>
<th>Perplexity (Evaluation test set 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1.00</td>
<td>608.08</td>
<td>637.25</td>
</tr>
<tr>
<td>Interpolated Model (LDA latent topic word count weighting)</td>
<td>0.10</td>
<td>443.96</td>
<td>467.60</td>
</tr>
<tr>
<td>Interpolated Model (Proposed weighting Scheme)</td>
<td>0.03</td>
<td>439.08</td>
<td>458.70</td>
</tr>
</tbody>
</table>

4.3. Error rate reduction

To evaluate the WER reduction, we used the WSJ1 development and evaluation test set 1. We have achieved significant WER reduction by our proposed weighting approach against the baseline model and the LDA latent topic word count weighting scheme. The results of the experiments are shown in Table 4. From the table we note that both weighting schemes outperform the baseline model. The proposed weighting approach achieves WER reductions of about 9.65% and 9.5% against the baseline model, about 1.48% and 1.00% against the latent topic word count weighting scheme for the development test set 1 and evaluation test set 1.

Table 4: WER results for WSJ1 Development and Evaluation test set1

<table>
<thead>
<tr>
<th>Language Model</th>
<th>WER(%) Development Test set 1</th>
<th>WER(%) Evaluation Test set 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>24.97</td>
<td>26.43</td>
</tr>
<tr>
<td>Interpolated Model (LDA latent topic word weighting)</td>
<td>22.90</td>
<td>24.16</td>
</tr>
<tr>
<td>Interpolated Model (Proposed Scheme)</td>
<td>22.56</td>
<td>23.92</td>
</tr>
</tbody>
</table>

5. Conclusions and future works

We proposed a novel weighting scheme for unsupervised language model adaptation using LDA. We compared two weighting schemes for topic model adaption. The LDA model is used for topic extraction and topic clusters are formed by a hard-clustering method. The proposed word count weighting of the topic generated by hard clustering outperforms the LDA latent topic word count weighting in both perplexity and WER measurement. We believe that LDA latent topic word count weighting is appropriate for unigram latent topic models extracted by LDA. We verified our idea on the Brown corpus and the WSJ1 training transcript text data.

For future work, we will use the proposed LM adaptation weighting approach in \( n \)-best list rescoring experiments. Moreover, we will apply the \( n \)-gram probabilities of the LDA adapted model as features in the minimum discriminant information (MDI) adaptation method. Also, we will employ the proposed weighting idea in HMM-LDA [19] too.

6. References