LDA-BASED LM ADAPTATION USING LATENT SEMANTIC MARGINALS AND MINIMUM DISCRIMINANT INFORMATION

Md. Akmal Haidar and Douglas O'Shaughnessy

INRS-EMT, 6900-800 de la Gauchetiere Ouest, Montreal (Quebec), H5A 1K6, Canada

ABSTRACT

We introduce an unsupervised language model (LM) adaptation approach using latent Dirichlet allocation (LDA) and latent semantic marginals (LSM). LSM is a unigram probability distribution over words and is estimated using the LDA model. A hard-clustering method is used to form topics. Each document is assigned to a topic based on the maximum number of words chosen from the topic for that document in LDA analysis. An LDA-adapted model is created using the weighted combination of topic models. The LDA-adapted model is modified by using the LSM as dynamic marginals, and a new adapted model is formed by using the minimum discriminant information (MDI) approach, which minimizes the distance between the new adapted model and the LDA-adapted model. We apply LM adaptation approaches for original and automatic (recognition results after first-pass decoding) transcriptions test data and have seen significant perplexity and word error rate (WER) reductions over a traditional approach.

Index Terms— Latent Dirichlet allocation, unsupervised LM adaptation, latent semantic marginals, speech recognition

1. INTRODUCTION

Speech recognition performance is reduced when the styles, domain or topics of the recognition tasks are different from the training set. The language model adaptation helps to exploit specific, albeit limited, knowledge about the recognition task to compensate for this mismatch [1].

Statistical n-gram models suffer from the lack of long-range information, which limits performance. So, it is important to handle long-range information. Various methods have been studied to extract the latent semantic information from a training corpus such as Latent Semantic Analysis (LSA) [2], Probabilistic Latent Semantic Analysis (PLSA) [3], and LDA [4]. In LSA, semantic information can be obtained from a word-document co-occurrence matrix. In PLSA and LDA, a set of probabilistic topics is introduced to show the semantic properties of words and documents. Here, a document is made out of a mixture of topics and a topic is a unigram probability distribution over words. The unigram topic models obtained by LDA are adapted to form LSM, which are used to modify a background model to form an adapted model with a constraint that the marginalized unigram probability distribution of the adapted model is equal to the LSM [5]. Here, we compare with the approach in [5] for the LDA-adapted topic model using the above marginals.

In this paper, we modified an LDA-adapted topic n-gram model [6] by using LSM as dynamic marginals, and find a final adapted model by using the minimum discriminant information (MDI), which uses KL divergence as the distance measure between probability distributions [7]. We employed LDA on the background corpus. A hard-clustering approach is used to form topic clusters. The weights of topic models are computed using the n-gram count of the topics generated by a hard-clustering method to form the LDA-adapted topic n-gram model [6]. A new adapted model is formed by minimizing the KL divergence between the new adapted model and the LDA-adapted topic n-gram model, subject to a constraint that the marginalized unigram distribution of the new adapted model is equal to the unigram distribution estimated by using the LDA model: this is called LSM [5]. Our approach is compared with a traditional approach used in the literature where an adapted model is formed by minimizing the KL divergence between the adapted model and the background model using the above constraint. The complete idea is illustrated in Figure 1. We used the adapted LM in the second pass of decoding. We applied the LM-adaptation approaches using both original and automatic test transcriptions data, and have seen that our approach gives significant reductions in perplexity and word error rate over the conventional approach [5].

The rest of this paper is organized as follows. In section 2, related works on language model adaptation using LDA and MDI are reviewed. Section 3 is used for reviewing the LDA model, topic clustering method and latent semantic marginals. LM adaptation methodology is described in section 4. In section 5, experiments and results are explained. Finally the conclusion is described in section 6.

2. RELATED WORK

Many methods have been studied to compensate for the limitations of n-gram models that capture only the local dependencies between words. An earlier approach is a cache-based

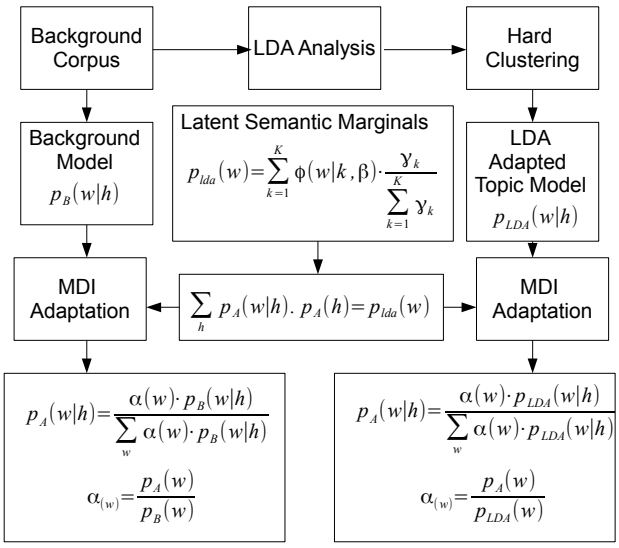

Fig. 1. LM adaptation using latent semantic marginals and MDI

language model that considers the idea that a word could occur again if it appeared earlier in a document. This helps to increase the probability of previously seen words in a document when predicting a future word [8]. The same idea was used in the trigger-based LM adaptation, which uses a maximum entropy approach [9] to raise the probability of unseen but topically related words.

Recently, latent topic analysis has been used broadly for language modeling. A hard-clustering approach is used to form topic clusters where a document is assigned by a single topic and used in LM adaptation [10]. However, LDA is one of the most popular probabilistic topic models and has been effectively used in recent research work in LM adaptation. The unigram topic models extracted by LDA are combined with a tri-gram baseline model, which achieved significant perplexity and WER reduction [11]. The topic clusters are formed by applying a hard-clustering approach on the document-topic matrix in LDA analysis. Perplexity reduction was shown by combining tri-gram topic LM's with the baseline tri-gram LM [12]. The unigram and n-gram counts of the topic generated by hard clustering are used to compute the mixture weights of the topic models and have shown significant improvement in perplexity and WER reductions [13, 6].

Various methods have been discussed previously for LM adaptation using MDI. The idea is to form an adapted model by modifying a background model with the minimization of the KL divergence between the background model and the adapted model. A constraint is induced that the marginalized unigram probability distribution of the adapted model is equal to an unigram distribution, which is estimated from some in-domain text data. The latter unigram distribution is called dynamic marginals [7, 14]. Here, an additional constraint is imposed to minimize the computational cost in computing the normalization term. The additional constraint is that the sum of the observed n-gram probabilities of the adapted model is equal to the sum of the observed n-gram probabilities of the background model. An LDA-adapted unigram distribution is used as the dynamic marginals instead of using a locally estimated unigram distribution [5]. Here, we modified the LDA-adapted topic trigram model using the latent semantic marginals described in [5], used the same additional constraint to minimize the computational cost, and have seen better performance over the other approach.

3. LDA, TOPIC CLUSTERING, AND LATENT SEMANTIC MARGINALS

3.1. Latent dirichlet allocation

LDA is a three-level hierarchical Bayesian model [4], where each item of a collection of discrete data is modeled as a finite mixture over an underlying set of topics. Then, each topic is 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 θ 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, θ indicates the relative importance of topics for a document and $\phi(w|k,\beta)$ represents the word probabilities conditioned on the topic k with a Dirichlet prior β and indicates the relative importance of particular words in a topic.

3.2. Topic clustering

We have used the MATLAB topic modeling toolbox [15] 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 occurrences of word w_j in topic z_k over the training set. In the DP matrix, an entry $DP(i,k)$ contains the total occurrences of words in document d_i that are from a topic $z_k (k = 1, 2, \ldots, K)$.

Topic clusters are formed by assigning a topic z_i^* to a document d_i as:

$$z_i^* = \arg\max_{1<k<K} DP(i,k) \quad (1)$$

i.e., a document is assigned to a topic from which it takes the maximum number of words. As a result, all the documents of the training corpus are assigned to K topics. Then K topic n-gram LM's are trained.

3.3. Latent semantic marginals

To compute the latent semantic marginals, we used the technique described in [5]. We first treat the test data (original or automatic transcription) as a single document. Then, we applied a Gibbs sampler for a new document to estimate the Dirichlet posterior over the topic mixture weights [15]. The LDA-adapted marginal is then computed as follows [5]:

$$p_{lda}(w) = \sum_{k=1}^{K} \phi(w|k,\beta).\frac{\gamma_k}{\sum_{k=1}^{K}\gamma_k}, \quad (2)$$

where γ_k is the weight of topic k for the test document obtained after LDA inference. $\phi(w|k,\beta)$ is the word probability for topic k obtained after applying LDA over the training set and is computed as [15]:

$$\phi(w|k,\beta) = \frac{WP(w,k)+\beta}{WP(.,k)+W\beta},$$

where $WP(.,k)$ is the total count of words in topic k and W is the total number of words. $WP(w,k)$ and β are defined as above.

4. LM ADAPTATION APPROACH

4.1. LDA-adapted topic mixture model generation

In the LDA model, 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:

$$p_{LDA}(w|h) = \sum_{i=1}^{K} \lambda_i p_{z_i}(w|h) \quad (3)$$

where $p_{z_i}(w|h)$ is the i^{th} topic model and λ_i is the i^{th} mixture weight. To find topic mixture weight λ_i, the n-gram count of the topics, generated by Equation 1, is used [6]. Therefore,

$$\lambda_k = \sum_{j=1}^{n} p(z_k|w_{j-n},\ldots,w_{j-1})p(w_{j-n},\ldots,w_{j-1}|d) \quad (4)$$

with

$$p(z_k|w_{j-n},\ldots,w_{j-1}) = \frac{TF(w_{j-n},\ldots,w_{j-1},k)}{\sum_{p=1}^{K} TF(w_{j-n},\ldots,w_{j-1},p)}$$

$$p(w_{j-n},\ldots,w_{j-1}|d) = \frac{freq(w_{j-n},\ldots,w_{j-1})}{\text{Total counts of all } n\text{-grams}}$$

where $TF(w_{j-n},\ldots,w_{j-1},p)$ represents the number of times the n-gram $(w_{j-n},\ldots,w_{j-1})$ is seen in topic p, which is created by Equation 1. $freq(w_{j-n},\ldots,w_{j-1})$ is the frequency of the n-gram $(w_{j-n},\ldots,w_{j-1})$ in document d.

4.2. Adaptation using latent semantic marginals

The goal of the LM adaptation using dynamic marginals [7] is to form an adapted model by minimizing the KL-divergence between the adapted model and the background model subject to the marginalization constraint for each word w in the vocabulary [5]:

$$\sum_{h} p_A(h).p_A(w|h) = p_{lda}(w). \quad (5)$$

The constraint optimization problem has close connection to the maximum entropy approach [9], which provides that the adapted model is a rescaled version of the background model:

$$p_A(w|h) = \frac{\alpha(w)}{Z(h)}.p_{B/LDA}(w|h)$$

with

$$Z(h) = \sum_{w} \alpha(w).p_{B/LDA}(w|h) \quad (6)$$

where $Z(h)$ is a normalization term, which guarantees that the total probability sums to unity, $p_{B/LDA}(w|h)$ is the background or LDA-adapted topic model, and $\alpha(w)$ is a scaling factor that is usually approximated as:

$$\alpha(w) \approx (\frac{p_A(w)}{p_{B/LDA}(w)})^{\delta}$$

where δ is a tuning factor between 0 and 1. In our experiments we used the value of δ as 0.5 [5]. We used the same procedure as [7] to compute the normalization term. To do this, an additional constraint is employed where the total probability of the observed transitions is unchanged:

$$\sum_{w:observed(h,w)} p_A(w|h) = \sum_{w:observed(h,w)} p_{B/LDA}(w|h).$$

The background and the LDA-adapted topic models have standard back-off structure and the above constraint, so the adapted LM has the following recursive formula:

$$p_A(w|h) = \begin{cases} \frac{\alpha(w)}{Z_o(h)}.p_{B/LDA}(w|h) & \text{if } (h,w) \text{ exists} \\ b(h).p_A(w|\hat{h}) & \text{otherwise} \end{cases}$$

where

$$Z_o(h) = \frac{\sum_{w:observed(h,w)} \alpha(w).p_{B/LDA}(w|h)}{\sum_{w:observed(h,w)} p_{B/LDA}(w|h)}$$

and

$$b(h) = \frac{1-\sum_{w:observed(h,w)} p_{B/LDA}(w|h)}{1-\sum_{w:observed(h,w)} p_A(w|\hat{h})}$$

where $b(h)$ is the back-off weight of the context h to ensure that $p_A(w|h)$ sums to unity. $\hat{h}$ is the reduced word history of h. The term $Z_o(h)$ is used to do normalization similar to Equation 6 except the summation is considered only on the observed alternative words with the equal word history h in the LM [5].

5. EXPERIMENTS AND RESULTS

5.1. Data and experimental setup

LM adaptation approaches are evaluated using the Wall Street Journal (WSJ1) corpus [16] transcription text data. We used all the training transcription text data (1,317,793 words) for training and development (7235 words) and the evaluation (6708 words) test set 1 for testing. Here, those sentences of the test set are kept where all the words of the sentences are in the dictionary. Since the transcripts used to train the LMs do not have any topic annotation, for the purpose of topic analysis, we split the training transcription text data into 300 sentences for each document and in total 261 documents are created. The total number of unique word tokens used for LDA analysis and LM generation is 20484.

We used the SRILM toolkit [17] and HTK toolkit [18] for our experiments. We trained LMs using the SRILM toolkit. The mixture weights are tuned on the development test set. We used perplexity and WER to measure the performance of the experiments. We used the baseline acoustic model from [19], where the model is trained by using all WSJ and TIMIT [20] training data, the 40 phones set of the CMU dictionary [21], 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 [19].

5.2. Perplexity reduction

LDA is applied on the WSJ1 training transcription text data to form 40 topic clusters. The trigram topic models are trained using the back-off version of Witten-Bell smoothing. The mixture weights of the topic models are computed using Equation 4. The LDA-adapted topic model is formed using Equation 3. We used latent semantic marginals (Equation 2) to adapt the background model and the LDA-adapted topic model subject to the constraint in Equation 5. We tested the LM-adaptation approaches for both original and automatic transcription test data. Automatic transcription is the recognition result obtained after first-pass decoding. The experimental results are described in Table 1 for original test transcription data and in Table 2 for automatic test transcription. All the adapted models give significant perplexity reductions over the background model. The language models in the third and fourth rows of Table 1 (original test transcription) show significant reduction in perplexity of about 2.18% and **45.76%** for the development test set and about 1.47% and **46.90%** for the evaluation test set, over the background model. For automatic test transcription, the language models in the third and fourth rows of Table 2 yield significant perplexity reductions of about 3.05% and **48.68%** for the development set and about 2.47% and **49.21%** for the evaluation set. However, the perplexities for automatic transcription test data are high because of recognition errors present after the first-pass decoding. For both original and automatic transcription test data, our approach outperforms the traditional approach used in the literature.

Table 1. *Perplexity results of the tri-gram model obtained by using LSM and MDI for original transcription*

Language Model	Perplexity-Development Set	Perplexity-Evaluation Set
Background	655.33	745.38
MDI adaptation of Background Model	640.99	734.36
MDI adaptation of LDA-adapted topic model	**355.39**	**395.74**

Table 2. *Perplexity results of the tri-gram model obtained by using LSM and MDI for automatic transcription*

Language Model	Perplexity-Development Set	Perplexity-Evaluation Set
Background	822.62	880.20
MDI adaptation of Background Model	797.50	858.43
MDI adaptation of LDA-adapted topic model	**422.11**	**447.02**

5.3. Word error rate reduction

We evaluated the WER experiments using lattice rescoring. In the first pass, we used the background language model for lattice generation. In the second pass, we applied the LM-adaptation approaches described in section 4 for lattice rescoring. We formed the automatic transcription by using first-pass decoding results. We applied the LM-adaptation approaches using both the original and automatic development test transcriptions. The experimental results are described in Table 3. From the table we can note that both the approaches outperform the background model. Moreover, our approach yields better results than the approach used in the literature. For the automatic transcription, we have seen improvements of about 0.23% and 0.27%, and about **2.30%** and **3.53%** WER reductions using the MDI adaptation of the background model and LDA-adapted topic model for the development

Table 3. *WER results for the WSJ1 Development and Evaluation test set1 using tri-gram models obtained by using LSM and MDI*

Language Model	WER (%): Development Set	WER (%): Evaluation Set
Background (First-Pass)	25.63	26.34
MDI adaptation using Automatic transcription:		
Background Model	25.57	26.27
LDA-adapted Topic Model	**25.04**	**25.41**
MDI adaptation using Original Transcription:		
Background Model	25.53	26.27
LDA-adapted Topic Model	**24.82**	**25.29**

and evaluation test sets respectively. As expected, the LM-adaptation using the original transcription test data performs better than using the automatic transcription obtained from the first pass.

6. CONCLUSIONS

An unsupervised language model adaptation approach using LDA, LSM and MDI is proposed. A hard-clustering approach is applied on the document-topic matrix in LDA analysis to form topic clusters. An n-gram weighting approach is used to compute the mixture weights of the component topic models. An LDA-adapted topic model is computed using the weighted combination of n-gram topic models. A new adapted model is formed by modifying the LDA-adapted topic model using MDI, which minimizes the KL divergence between the new adapted model and the LDA-adapted topic model subject to a constraint that the marginalized unigram probability distribution of the new adapted model is equal to a unigram probability distribution estimated by using the LDA model, called LSM. We used LM adaptation approaches in the second pass of decoding. We verified our approach for both the original and automatic transcriptions. A traditional MDI adaptation of a background model using the same constraint is compared with our approach. We have seen that our approach gives significant reductions in perplexity and WER over the traditional approach used in the literature.

7. REFERENCES

[1] J. R. Bellegarda, "Statistical Language Model Adaptation: Review and perspective", *Speech Communication*, vol. 42, pp. 93-108, 2004.

[2] J. R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling", in *IEEE Trans. on Speech and Audio Proc*, vol. 88, No. 8, pp. 1279-1296, 2000.

[3] D. Gildea and T. Hofmann, "Topic-Based Language Models Using EM", in *Proc. of EUROSPEECH*, pp. 2167-2170, 1999.

[4] D. M. Blei, A. Y.Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[5] Y.-C. Tam and T. Schultz, "Unsupervised Language Model Adaptation Using Latent Semantic Marginals", in *Proc. of INTERSPEECH*, pp. 2206-2209, 2006.

[6] M. A. Haidar and D. O'Shaughnessy, "Unsupervised Language Model Adaptation Using N-gram weighting", in *Proc. of CCECE*, pp. 857-860, 2011.

[7] R. Kneser, J. Peters, and D. Klakow, "Language Model Adaptation Using Dynamic Marginals", in *Proc. of EUROSPEECH*, pp. 1971-1974, 1997.

[8] R. Kuhn and R. D. Mori, "A Cache-Based Natural Language Model for Speech Recognition", in *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 12(6), pp. 570-583, 1990.

[9] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling", *Computer, Speech and Language*, vol. 10(3), pp. 187-228, 1996.

[10] R. Iyer and M. Ostendorf, "Modeling Long Distance Dependence in Language: Topic Mixtures vs Dynamic Cache Models", in *Proc. of ICSLP*, vol. 1, pp. 236-239, 1996.

[11] Y.-C. Tam and T. Schultz, "Dynamic Language Model Adaptation Using Variational Bayes Inference", in *Proc. of INTERSPEECH*, pp. 5-8, 2005.

[12] F. Liu and Y. Liu, "Unsupervised Language Model Adaptation Incorporating Named Entity Information", in *Proc. of ACL*, pp. 672-679, 2007.

[13] M. A. Haidar and D. O'Shaughnessy, "Novel Weighting Scheme for Unsupervised Language Model Adaptation Using Latent Dirichlet Allocation", in *Proc. of INTERSPEECH*, pp. 2438-2441, 2010.

[14] M. A. Haidar and D. O'Shaughnessy, "Unsupervised Language Model Adaptation Using Latent Dirichlet Allocation and Dynamic Marginals", in *Proc .of EUSIPCO*, pp. 1480-1484, 2011.

[15] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics", in *Proc. National Academy of Sciences*, 101 (suppl. 1), pp. 5228-5235, 2004.

[16] "CSR-II (WSJ1) Complete", Linguistic Data Consortium, Philadelphia, 1994.

[17] A. Stolcke, "SRILM- An Extensible Language Modeling Toolkit", in *Proc. of ICSLP*, vol. 2, pp. 901-904, 2002.

[18] S. Young, P. Woodland, G. Evermann and M. Gales, "The HTK toolkit 3.4.1", http://htk.eng.cam.ac.uk/, Cambridge Univ. Eng. Dept. CUED.

[19] K. Vertanen, "HTK Wall Street Journal Training Recipe", http://www.inference.phy.cam.ac.uk/kv227/htk/.

[20] John S. Garofolo, et al,"TIMIT Acoustic-Phonetic Continuous Speech Corpus" Linguistic Data Consortium, Philadelphia, 1993

[21] "The Carnegie Mellon University (CMU) Pronounciation Dictionary",http://www.speech.cs.cmu.edu/cgi-bin/cmudict