Language Model Adaptation Based on Topic Probability of Latent Dirichlet Allocation

Hyung-Bae Jeon and Soo-Young Lee

Two new methods are proposed for an unsupervised adaptation of a language model (LM) with a single sentence for automatic transcription tasks. At the training phase, training documents are clustered by a method known as Latent Dirichlet allocation (LDA), and then a domain-specific LM is trained for each cluster. At the test phase, an adapted LM is presented as a linear mixture of the now trained domain-specific LMs. Unlike previous adaptation methods, the proposed methods fully utilize a trained LDA model for the estimation of weight values, which are then to be assigned to the now trained domain-specific LMs; therefore, the clustering and weight-estimation algorithms of the trained LDA model are reliable. For the continuous speech recognition benchmark tests, the proposed methods outperform other unsupervised LM adaptation methods based on latent semantic analysis, non-negative matrix factorization, and LDA with *n*-gram counting.

Keywords: Language model adaptation, topic model, Latent Dirichlet allocation, weighted mixture model, LDA.

I. Introduction

To compensate for mismatches in the domain, topic, or style between training and test tasks, a language model (LM) adaptation is required to be introduced during a speech recognition test.

Various LM adaptation techniques have been proposed in the literature, and can be categorized into two approaches, model interpolation and constrained adaptation [1].

In an interpolation-based approach, a test-domain corpus is used to train a domain-specific LM, which is later combined with a baseline LM obtained from a training corpus. A simple weighted sum of two word probabilities is used for an interpolation.

In a constrained-adaptation approach, a baseline LM is adapted with constraints obtained from a test-domain corpus. In the popular minimum discrimination information (MDI) method, a baseline LM is adapted to minimize a Kullback–Leibler divergence between probability distributions of the adapted LM and the baseline LM, while satisfying the constraints of matched unigram distributions in the baseline and test corpora [2], [3].

Although constrained-adaptation approaches are potentially more advantageous than interpolation approaches, they also require a bigger test corpus.

For many real-world automatic transcription applications, only one sentence or a few paragraphs may be available for a test. In this case, a domain-specific training of an LM is not feasible, and only a simple mixture model of relevant domain LMs is applicable. In such an approach, a baseline LM would need to be pre-trained with a training corpus as a linear mixture of domain LMs [4], [5]. In addition, in such a case, only the weights associated with the domain LMs would need to be

Manuscript received May 30, 2015; revised Feb. 1, 2016; accepted Feb. 29, 2016.

This work was supported by the ICT R&D program of MSIP/IITP (R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning).

Hyung-Bae Jeon (corresponding author, hbjeon@etri.re.kr) is with the Department of Bio and Brain Engineering, KAIST, and also with the SW & Contents Research Laboratory, ETRI, Deajeon, Rep. of Korea.

Soo-Young Lee (sylee@kaist.ac.kr) is with the Department of Electrical Engineering and Department of Bio and Brain Engineering, KAIST, Daejeon, Rep. of Korea.

estimated from the very small amount of available test data. The domains (or topics) would usually be learned by unsupervised document clustering algorithms, and each domain LM would be trained with documents from a specific domain only; latent semantic analysis (LSA) [6], probabilistic LSA [7], and non-negative matrix factorization (NMF) [8] are used for the clustering.

Recently, Latent Dirichlet allocation (LDA) has become popular for unsupervised topic modeling with a controllable sparseness of latent topic distribution [9], [10]. A linear mixture model may be further adapted by model interpolation or constrained-adaptation algorithms; for example, a latent semantic marginal (LSM) method using LDA-adapted unigrams with related MDI constraint [11], [12]. In the case of LM adaptations, many speech recognition applications can use only an automatic transcription sentence generated from a first-pass decoding. For this reason, a method incorporating a mixture of domain LMs is more appropriate for an LM adaptation since we only need find an efficient method for accurately estimating the weights associated with the domain LMs. LDA has outstanding performance with regard to topic modeling [9], [10], which leads to a robust estimation of the weights associated with the domain LMs. In the case of an LM adaptation based upon a linear mixture of domain LMs, we propose two new methods to estimate, from a single test sentence, the weights associated with these domain LMs.

The proposed methods are based on an LDA topic model. Unlike previous LDA-based methods [1], the proposed methods fully utilize both a learned topic and word probabilities for a given domain-LM weight estimation. The proposed methods for LM adaptation perform equally or better when compared with other approaches in the literature (namely, those that are based on one of three clustering algorithms: LSA, NMF, or LDA).

The rest of this paper is organized as follows. In Section II, the proposed methods for LM adaptation are presented, and the details of our experiments are described in Section III. Finally, conclusions are drawn in Section IV.

II. LDA-Based LM Adaptation

1. Mixture Model for LM Adaptation

Figure 1 shows a flow diagram illustrating the process of an LM adaptation and lattice rescoring in a test phase suitable for an automatic transcription application. For each test dialogue, the "First-Pass Decoding" module generates a "word lattice" and an "automatic transcription" from a baseline speech recognition system. The transcription results may be improved by adapting the LM for each test dialogue.

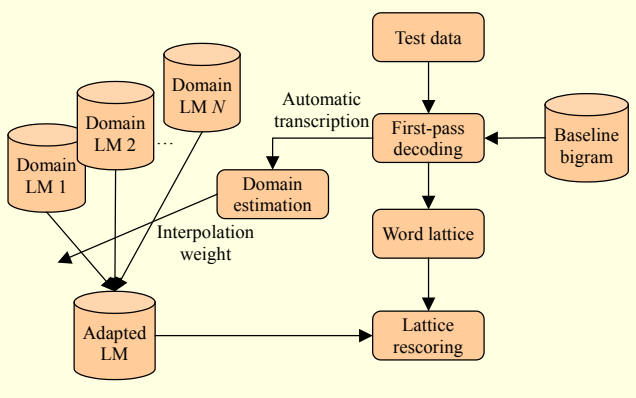

Fig. 1. Flow diagram of proposed LM adaptation.

In this paper, we propose two methods for estimating accurate interpolation weights with respect to the particular type of LM adaptation illustrated in Fig. 1.

An n-gram LM defines a probability distribution over sequences of words. Given a word sequence, $w_1, \ldots, w_J$, an LM approximates its associated probability, $p_A(w_1, \ldots, w_J)$, using the following equation:

$$p_A(w_1, \ldots, w_J) = \prod_{j=1}^{J} p_A(w_j \mid h_j), \quad (1)$$

where $(w_1, \ldots, w_J)$ represents a J-word test dialogue, and h_j denotes the history (previous words) at time j.

For an n-gram model, a Markovian assumption results in an $(n-1)$ word sequence for h_j and (1) holds only approximately. The goal of the LM adaptation depicted in Fig. 1 is to provide a better estimate for the probability given in (1).

In the case of a weighted mixture model, the n-gram probability of the adapted LM depicted in Fig. 1 is represented as

$$p_A(w_j \mid h_j) = \sum_{k=1}^{K} \lambda_k p_{D_k}(w_j \mid h_j), \quad (2)$$

where $p_{D_k}(w_j \mid h_j)$ refers to the n-gram probability in the domain D_k; λ_k is the weight of the kth domain D_k; and K is the number of domains or topics in the training corpus. Therefore, provided that domain-specific LMs are trained in advance, the LM adaptation renders a task to find proper values for weight λ_k from a given first-pass transcription result.

2. Unsupervised Document Clustering for Domain LMs

The domain-specific LMs are trained in advance with portions of the training corpus, which are clustered by unsupervised learning algorithms. Following a "bag-of-words" model, the input to the clustering algorithms is determined as a normalized word-document matrix, **W**, of which the (m, l)th

element is [6]

$$\mathbf{W}_{ml} = (1-\varepsilon_m)\frac{c_{ml}}{n_l}, \quad (3)$$

where c_{ml} is the number of times word w_m occurs in document d_l, n_l is the total number of words present in document d_l, and ε_m is the normalized entropy of word w_m in the corpus. Here, M and L are the number of words in the vocabulary and the number of documents, respectively. The normalized entropy ε_m of the mth word is introduced to place more emphasis on discriminant words, and is defined as

$$\varepsilon_m = -\frac{1}{\log L}\sum_{l=1}^{L}\frac{c_{ml}}{\sum_l c_{ml}}\log\frac{c_{ml}}{\sum_l c_{ml}}. \quad (4)$$

The training corpus may be clustered by LSA, NMF, and LDA. LDA is especially interesting with both its three-level hierarchical architecture and its capability to represent documents by random latent topics.

Figure 2 shows a graphical model representation of LDA. In LDA, latent topics are characterized by a distribution over words. LDA learns parameters α, θ, Z, β, and η by maximizing the likelihood of the word-document matrix $\mathbf{W}$ obtained from the training corpus. The random variable Z_{ml} is a word-level topic assignment that assigns $\mathbf{W}_{ml}$ to one of K topics. The variable Z_{ml} follows a multinomial distribution with parameter θ_l, which represents topic probabilities of the lth document over the training corpora. The random variable β_k represents word probabilities of the kth topic. The parameters α and η control Dirichlet distributions of θ_l and β_k, respectively. The parameter learning is performed by variational inference [9] or by Gibbs sampling [12].

Once the LDA parameters are learned, the clustering may be conducted from $\theta_l = [\theta_{l1}, \theta_{lK}]$; that is, the topic probability of a document. The popular choice is to find the topic k_l that gives the maximum θ_{lk} for the lth document as [12]

$$k_l = \arg\max_k \theta_{lk}, \qquad 1 \le k \le K. \quad (5)$$

In our experiments, we used up to three topics for each document. This multi-cluster assignment scheme allows overlap among clusters with smooth transition and enriches the training document of each cluster. In addition, a k-means clustering algorithm may be used with a Hellinger distance measure between the lth and tth documents as [10]

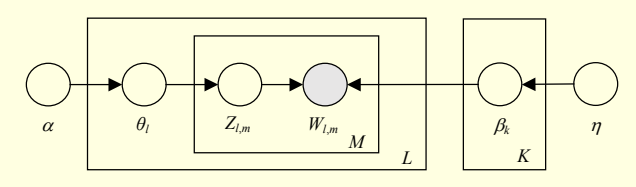

Fig. 2. Graphical representation of LDA.

$$D_{lt} = \sum_{k=1}^{K}\left(\sqrt{\theta_{lk}} - \sqrt{\theta_{tk}}\right)^2. \quad (6)$$

Other distance measures, such as Euclidean distance, may also be used without any significant differences occurring. In this case, the number of clusters may be different from the number of topics K. Accordingly, documents that have a similar topic probability are assigned to the same domain. Then, n-gram domain (or topic-specific) LMs are trained from the documents in the same clusters.

3. LM Adaptation Based on Domain Estimation

In contrast to a fixed coefficient method, more appropriate clusters were chosen in the case of our given automatic transcription [13]. Furthermore, the weights associated with the domain LMs were estimated by the topic probability.

For any given test dialogue, a first-pass decoding provides a transcription estimate, of which the associated word probability vector is equivalent to a column from the word-document matrix $\mathbf{W}$. Therefore, it is possible to obtain a topic-probability vector $\boldsymbol{\theta}_t$ of the test dialogue from the trained LDA parameters. Then, weight parameters ($\lambda_k, 1 \le k \le K$) are estimated to give the n-gram probability in (2).

In [12], the weight λ_k was estimated from an n-gram count of the topic as

$$\lambda_k = \sum_{i=1}^{I} p(k \mid \boldsymbol{w}_i)\, p(\boldsymbol{w}_i \mid D_t), \quad (7)$$

where $\boldsymbol{w}_i$ is the ith n-gram and I is the number of n-grams in the test document D_t. This method will be denoted by LDA-NC (n-gram count) hereafter.

For a short test dialogue, only a very small fraction of all the bigrams or trigrams is available, and the estimated weights (of the domain LMs) may not be robust. To overcome this difficulty, two new methods are proposed for the said weight estimation. Both methods utilize the topic probability vector $\boldsymbol{\theta}_t$ evaluated for the test dialogue from the LDA parameters. Therefore, both clustering and weight estimation are based on the same LDA parameters, and consistency is maintained.

First, we use the topic probability with normalization as

$$\lambda_k = \frac{\boldsymbol{\theta}_{tk}}{\sum_k^T \boldsymbol{\theta}_{tk}}, \quad (8)$$

where T ($\le K$) is the number of topic mixtures to be considered. The second method utilizes the distance measure in (6); thus, the mixture weight becomes

$$\lambda_k \propto \left(\frac{D_{tk}}{\sum_k^T D_{tk}}\right)^{-1}, \quad (9)$$

where D_{tk} is the distance between the test vector $\boldsymbol{\theta}_t$ and average of all documents in the kth domain, θ_k. Here, T is the number of domains over all the training corpora. The proposed methods in (8) and (9) will be denoted by LDA-TP (topic probability) and LDA-DD (domain distance), respectively.

III. Experimental Results.

1. Experimental Setup

To show robustness on the training data, different speech corpora were used for the training module. The Hidden Markov Model Toolkit was used to train a hidden Markov model (HMM)-based acoustic model, first-pass decoding, lattice generation, and lattice rescoring. Three speech corpora, the WSJ1 SI-284, WSJ0, and TIMIT training sets, were used for training 7,238 state-tying HMMs. A given feature vector consists of one log energy coefficient and 12 Mel-frequency cepstral coefficients, and their first and second derivatives.

For the clustering and training of baseline LMs, four different corpora were used: (a) randomly selected 300K documents from 490K documents in the LDC CSR3 training text corpus from newswires between 1987 and 1994; (b) randomly selected 100K documents from 125K training documents in the LDC HUB4 text corpus from broadcast news between 1992 and 1996; (c) the full 4,876 telephone conversational dialogue in the part 1 transcription data of the LDC switchboard; and (d) 200K social network documents composed of 9M randomly selected tweets out of 54M tweets from the Twitter service. The total size of the LM training corpora was about 2.1 GB. These 604K documents were clustered by LDA, LSA, and NMF for a performance comparison. Then, the domain LMs were trained with Kneser-Ney discounting and entropy-based pruning algorithms by the SRI LM toolkit [14].

The Wall Street Journal (WSJ) corpus was used to evaluate the performance of the proposed methods for LM adaption in continuous speech recognition (CSR) tasks. Specifically, the November 1992 and November 1993 ARPA CSR benchmark test data (WSJ Nov 92 and WSJ Nov 93) were used. LM adaptation was achieved by (2), of which weight λ_k was estimated by (8) and (9) for the proposed LDA-TP and LDA-DD, respectively. Four other LM adaptation methods in the literature (LSA, NMF, LSM, and LDA-NC) were also implemented for a performance comparison.

2. Experimental Results

In Fig. 3, the distribution of topic probability θ_{lk} is plotted for the training and test corpora after LDA training with 10 topics ($K = 10$). For each document, the θ_{lk} values were sorted,

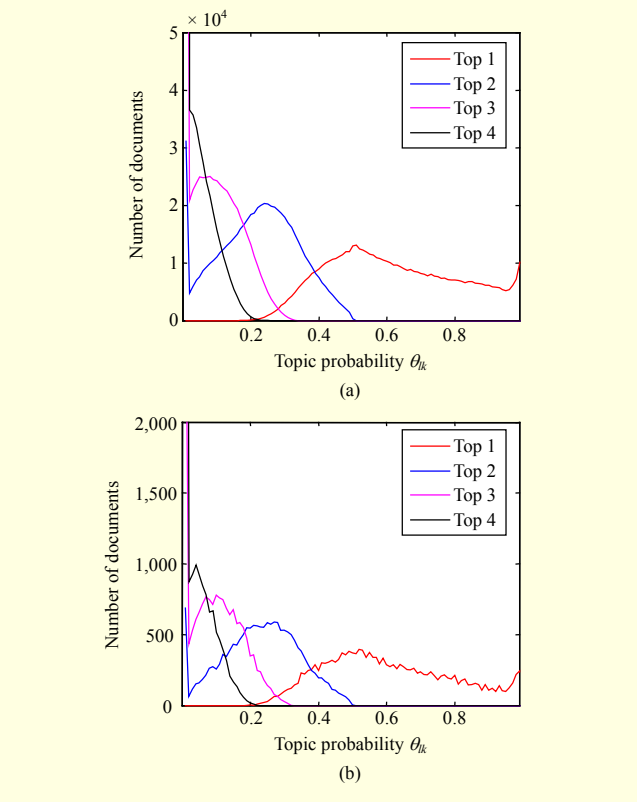

Fig. 3. Distribution of topic probability θ_{lk} values for (a) training corpora and (b) test corpora. For many documents only few θ_{lk} have non-zero values, and solid red line is quite distinct.

and the top four values were plotted. For example, the solid red line is a histogram depicting the largest values of θ_{lk}, and the blue solid line is a histogram depicting the second-largest values of θ_{lk}. It is clear that the solid red line is quite distinct; that is, it contains only a few non-zero values. Actually, the topic probability θ_{lk} is sparse, and the kurtosis value is 5.96 for the test corpora, which shows that LDA works well for unsupervised clustering. Further, the distribution does not show much difference between the training and test corpora, and the clustering is well generalized.

Figure 4 shows the number of documents assigned to each topic by (5); that is, assigned to the topic with the highest θ_{lk} value. The majority of the switchboard corpus and a big portion of the HUB4 corpus consist of one topic, while Twitter data from social networks form another topic. The CSR3 corpus is quite general and scattered across many other topics.

The performances of the various LM adaptation methods is evaluated under automatic transcription. Table 1 summarizes the perplexity improvements of the LM adaptation methods at the sentence level.

The performance of the proposed LDA-TP and LDA-DD methods is compared with that of existing methods, LSA,

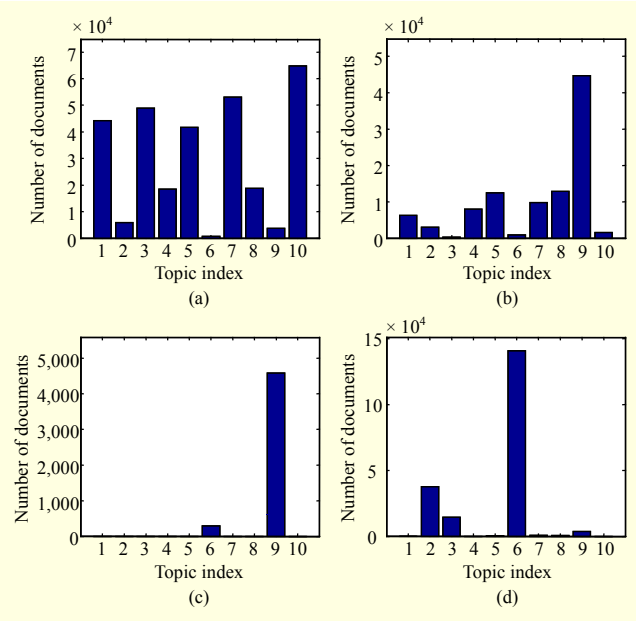

Fig. 4. Number of assigned documents to 10 topics by LDA: (a) LDC CSR3 corpus, (b) LDC HUB4 corpus, (c) LDC switchboard corpus, and (d) Twitter data.

Table 1. Perplexity values of various LM adaptation methods for WSJ Nov 92 and WSJ Nov 93 corpora.

LM adaptation	No. of mixtures for adaptation	Without final interpolation		With final interpolation		Average
		Nov 92	Nov 93	Nov 92	Nov 93	
Baseline LM	N/A	214.95	213.04	N/A	N/A	214.00
LSA	3	166.16	166.18	166.51	166.84	166.42
	10	167.61	168.59	169.10	169.86	168.79
NMF	3	167.10	176.78	165.89	172.95	170.68
	10	169.00	170.93	170.34	172.01	170.57
LSM	N/A	186.24	179.88	179.35	182.47	181.98
LDA-NC (*n*-gram)	3	165.90	161.39	167.99	163.59	164.67
	10	168.03	165.11	169.49	166.57	167.30
Proposed LDA-DD	3	**157.17**	**161.09**	**158.55**	**161.33**	**159.54**
	10	171.17	172.54	172.49	173.59	172.45
Proposed LDA-TP	3	**152.04**	**150.89**	**153.22**	**152.32**	**152.12**
	10	**151.90**	**150.93**	**153.61**	**152.65**	**152.27**

NMF, LSM, and LDA-NC. The best results from the final interpolation with the baseline LM were chosen out of ten different interpolation parameters. Except for the LDA-DD, the number of topics is set to ten (K = 10) for all cases. For the LDA- DD, 50 topics are initially learned by the LDA, but the latter ten clusters were formed by a k-means algorithm with

Table 2. Word error rate (%) of various LM adaptation methods for WSJ Nov 92 and WSJ Nov 93 corpora.

LM adaptation	No. of mixtures for adaptation	Without final interpolation		With final interpolation		Average
		Nov 92	Nov 93	Nov 92	Nov 93	
Baseline LM	N/A	10.83	13.17	N/A	N/A	12.00
LSA	3	10.12	12.77	10.15	12.80	11.46
	10	**9.69**	**12.25**	**9.69**	12.30	**10.98**
NMF	3	9.96	13.12	9.92	12.65	11.41
	10	9.80	12.45	9.84	12.36	11.11
LSM	N/A	10.70	13.32	10.17	13.03	11.81
LDA-NC (*n*-gram)	3	10.21	12.33	10.07	12.19	11.20
	10	10.03	**12.22**	9.96	**12.07**	11.07
Proposed LDA-DD	3	9.85	12.59	9.78	12.68	11.23
	10	9.87	12.30	9.94	12.30	11.10
Proposed LDA-TP	3	**9.66**	12.36	**9.66**	**12.19**	**10.97**
	10	**9.64**	**12.10**	**9.64**	**12.10**	**10.87**

distance measure (6). The number of non-zero weights (associated with the domain LMs) in (2) was set to three or ten.

Among the several LM adaptation methods, the performance of the proposed LDA-TP forms a distinct top group with the best (bold fonts) perplexity. In this top group, the perplexity was reduced from 214 to about 150 on average. The proposed LDA-DD, with up to three mixtures, also worked well and resulted in the second best (bold italic fonts) performance. The final interpolation with the baseline LM had the same form as the weighted mixture models and did not improve the performance much.

Table 2 summarizes the word error rates (WERs) of the various LM adaptation methods. For the WERs, the LSA, LDA-NC, and LDA-TP form a top group, and reduce WERs from 12% to less than 11%. The LSA and LDA-NC did poorly in terms of the perplexity measure. However, the proposed LDA-TP belongs to the top group for both perplexity and WERs measures, and for both the WSJ Nov 92 and Nov 93 corpora.

In the case of real-world applications, a smaller number of topic mixtures is necessary for a low computational complexity. In such a case, the proposed LDA-TP would have the best performance; that is, because the topic probability of the LDA is sparse, the proposed LDA-TP has a low level of degradation.

In Table 3, WERs are shown for several different numbers of topics for the proposed LDA-TP method. For all cases, the same number is used for both the LDA topics and LM clusters. Compared to Tables 1 and 2, a slightly better performance was obtained with more than ten topics. However, WERs were not

Table 3. Low sensitivity of word error rate (%) on number of topics and clusters for proposed LDA-TP method.

No. of topics/mixture	Perplexity		WER (%)	
	Nov 92	Nov 93	Nov 92	Nov 93
Baseline LM	214.95	213.04	10.83	13.17
5	171.43	168.62	10.07	12.94
10	**152.04**	**150.89**	**9.66**	**12.36**
30	133.37	131.75	**9.60**	**12.30**
50	**131.56**	**130.09**	10.07	12.86

Table 4. Robustness of word error rate (%) on adaptation data for proposed LDA-TP method.

Adaptation data	Perplexity		WER (%)	
	Nov 92	Nov 93	Nov 92	Nov 93
Test document	149.54	147.58	9.84	12.19
Test sentence	151.90	150.93	9.66	12.25

sensitive to the number of topics and clusters.

Table 4 shows the perplexity and WERs for different adaptation data. In general, the document-based estimation of the weights associated with the domain LMs shows a better performance than the sentence-based estimation. However, in the case of real-world applications, the adaptation data are restricted to an automatic transcription, and in Table 4 the sentence-based estimation results in a similar performance as the document-based estimation. Because the proposed LDA-TP method estimates the probabilities of the basis components (topics) that comprise a given domain, it gives a more robust performance in terms of the amount of adaptation data used.

IV. Conclusion

In this paper, we presented new methods for an LM adaptation based on the topic probability of LDA. Even with one test sentence, the proposed LDA-TP resulted in an excellent performance in terms of both perplexity and WER measures. In addition, the performance is robust to the number of clusters. This excellent performance may come from the consistency in the clustering and weight-estimation methods, both of which are based on LDA.

References

[1] J.R. Bellegarda, "Statistical LM Adaptation: Review and Perspectives," *Speech Commun.*, vol. 42, no. 1, Jan. 2004, pp. 93–108.

[2] R. Kneser, J. Peters, and D. Klakow, "Language Model Adaptation Using Dynamic Marginals," *European Conf. Speech Commun. Technol.*, Rhodes, Greece, Sept. 1997, pp. 1971–1974.

[3] M. Federico, "Efficient Language Model Adaptation through MDI Estimation," *European Conf. Speech Commun. Technol.*, Budapest, Hungary, Sept. 1999, pp. 1583–1586.

[4] Y. Si et al., "Block-Based Language Model for Target Domain Adaptation towards Web Corpus," *J. Computational Inf. Syst.*, vol. 9, Nov. 2013, pp. 9139–9146.

[5] K. Thadani, F. Biadsy, and D.M. Bikel, "On-the-fly Topic Adaptation for YouTube Video Transcription," *Interspeech*, Portland, OR, USA, Sept. 2012. pp. 210–213

[6] J.R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling," *Proc. IEEE*, vol. 88, no. 8, 2000, pp. 1279–1296.

[7] Y. Akita and T. Kawahara, "Language Model Adaptation Based on PLSA of Topics and Speakers for Automatic Transcription of Panel Discussions," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 3, Mar. 2005, pp. 439–445.

[8] W. Xu, X. Liu, and Y. Gong, "Document Clustering Based On Non-negative Matrix Factorization," *Ann. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval,* Toronto, Canada, July 2003, pp. 267–273.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Res.*, vol. 3, Feb. 2003, pp. 993–1022.

[10] D.M. Blei and J.D. Lafferty, "TOPIC MODELS," *Text Mining: Classification, Clustering, and Applications*, vol. 10, no. 71, June 2009, p. 34.

[11] Y.C. Tam and T. Schultz, "Unsupervised Language Model Adaptation Using Latent Semantic Marginals," *Interspeech*, Pittsburgh, PA, USA, Sept. 2006, pp. 2206–2209.

[12] M.A. Haidar and D. O'Shaughnessy, "Unsupervised LM Adaptation Using LDA-Based Mixture Models and Latent Semantic Marginals," *Comput. Speech Language*, vol. 29, no. 1, 2015, pp. 20–31.

[13] L. Xiaoyong and W.B. Croft, "Cluster-Based Retrieval Using Language Model," *Ann. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Sheffield, UK, July 2004, pp. 186–193.

[14] A. Stolcke, "Entropy-Based Pruning of Backoff Language Models," *DARPA Broadcast News Transcription Understanding Workshop*, Lansdowne, PA, USA, Feb. 1998, pp. 270–274.

Hyung-Bae Jeon received his BS degree in electronics engineering from Yonsei University, Seoul, Rep. of Korea, in 1999 and his MS degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2001. Since 2001, he has been with the Spoken Language Processing Research Section of ETRI, where he is now a principal researcher. His main research interests include speech recognition, language modeling, and machine learning.

Soo-Young Lee received his BS degree in electronics from Seoul National University, Rep. of Korea, in 1975; his MS degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Rep. of Korea, in 1977; and his PhD degree in electrophysics from the Polytechnic Institute of New York, USA in 1984. In early 1986, he joined the Department of Electrical Engineering, KAIST, as an assistant professor and is now a full professor with the Department of Electrical Engineering and also the Department of Bio & Brain Engineering. From June 2008 to June 2009, he also worked for the Mathematical Neuroscience Laboratory, RIKEN Brain Science Institute, Saitama, Japan, as part of his sabbatical leave. In 2000, he was the president of the Asia-Pacific Neural Network Assembly. Currently, he is president-elect of the newly-founded Asia-Pacific Neural Network Society. His research interests include the artificial brain, alias artificial cognitive systems, and human-like intelligent systems/robots based on biological information processing mechanisms in our brain. He has worked on computational models of the auditory and visual pathways; unsupervised and supervised learning architectures and algorithms; active learning; situation awareness from environmental sound; and top-down selective attention.