Language Model Adaptation Using Dirichlet Class Language Model Based on Part-of-Speech

Ali Hatami*
Computer Engineering Department, Iran University of Science and Technology, Tehran, Iran
ali_hatami@comp.iust.ac.ir
Ahmad Akbari
Computer Engineering Department, Iran University of Science and Technology, Tehran, Iran
akbari@iust.ac.ir
Babak Nasersharif
Electrical and Computer Engineering Department, K. N. Toosi University of Technology, Tehran, Iran
bnasersharif@kntu.ac.ir

Received: 27/Jul/2013 Accepted: 13/Jan/2014

Abstract

Language modeling has many applications in a large variety of domains. Performance of this model depends on its adaptation to a particular style of data. Accordingly, adaptation methods endeavour to apply syntactic and semantic characteristics of the language for language modeling. The previous adaptation methods such as family of Dirichlet class language model (DCLM) extract class of history words. These methods due to lake of syntactic information are not suitable for high morphology languages such as Farsi. In this paper, we present an idea for using syntactic information such as part-of-speech (POS) in DCLM for combining with one of the language models of n-gram family. In our work, word clustering is based on POS of previous words and history words in DCLM. The performance of language models are evaluated on BijanKhan corpus using a hidden Markov model based ASR system. The results show that use of POS information along with history words and class of history words improves performance of language model, and decreases the perplexity on our corpus. Exploiting POS information along with DCLM, the word error rate of the ASR system decreases by 1.2% compared to DCLM.

Keywords: Speech Recognition, Language Model Adaptation, Part-of-Speech, Perplexity, Word Error Rate.

1. Introduction

Statistical language modeling (SLM) has been successfully applied to many natural language and speech processing. The purpose of LM is to assign probabilities to sequences of words according to a certain distribution. Speech recognition focuses on searching for the best word sequence $\widehat{W}$ by maximizing a posteriori (MAP) probability of speech utterance X [1]:

$$\widehat{W} = \operatorname{argmax} p(W|X) = \operatorname{argmax} p(X|W)\, p(W) \qquad (1)$$

where p(X|W) is the acoustic likelihood given the hidden Markov model (HMM), and p(W) is the prior word probability given the LM. N-gram LM is a known approach that assigns probability to next word based on its immediately preceding n-1 history words. In an n-gram model [2], the probability of a word sequence ($w_1^T = (w_1, \dots, w_T)$) is calculated by multiplying the probabilities of predicted word w_i conditioned on its preceding $n-1$ words depicted by w_{i-n+1}^{i-1}:

$$p(W) = \prod_{i=1}^{T} p\left(w_i|w_1^{i-1}\right) \cong \prod_{i=1}^{T} p\left(w_i|w_{i-n+1}^{i-1}\right) \qquad (2)$$

where $p\left(w_i|w_1^{i-1}\right)$ shows the conditional probability of w_i given w_1^{i-1}. Factored language model (FLM) [3] is another kind of n-gram models. The FLM was proposed using factors for each word. In a FLM, a word is considered as a vector of K factors.

$$w_t = \{f_t^1, \dots, f_t^K\} \qquad (3)$$

These factors can be anything, including morphological classes, stems, roots and other such features.

The n-gram models suffer from the insufficiencies of long-distance information, which limit the model performance. To compensate this, n-gram model can be combined with the adaptation methods like latent Dirichlet allocation (LDA) that extract the semantic information. LDA [4] provides a powerful mechanism for discovering the structure of a text document. The latent topic of each document is treated as a random variable. To tackle the data sparseness and extract the large-span information for n-gram models, in [5], a new Dirichlet class LM (DCLM) is constructed. In this technique, the latent variable reflects the class of an n-gram event rather than the topic in LDA model. In addition, Cache DCLM (CDCLM) [5] is proposed to improve DCLM by considering dynamic classes of history words in the online estimation.

The previous adaptation methods just used semantic information and did not consider syntactic features. In the languages with high morphology such as Farsi, exploiting the syntactic information such as POS can be useful.

In this paper, we proposed a technique for using POS along with adaptation methods to improve language

* Corresponding Author

model and so speech recognition rate. In our DCLM based approach, word clustering is performed exploiting previous POS and history words. In this technique, we calculate the word probability given POS of previous words along with history words.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of related works. In Section 3, we discuss the proposed technique for using POS along with adaptation methods. Section 4 evaluates the LMs performance. Finally, in Section 5 we conclude our paper.

2. Related Work

2.1 Latent Dirichlet Allocation

Topic-based model is a common method for extracting semantic information from text corpus in order to adapt a language model. In the last decade, a variety of probability topic modeling approaches has been proposed to analyze the latent topics and meaning of documents and words, such as latent Dirichlet allocation (LDA). Blei et al. [4] introduced LDA by incorporating the Dirichlet priors for extracting the topic structure of a document. LDA builds a hierarchical Bayesian model. In this model, documents are represents by the random latent topics, which are specified by the distributions over words. In other words, LDA discovered the topic at document level and were used for building topic-based language model [6].

LDA was shown effective in document classification [4] and speech recognition [5]. LDA model defines two parameters consist of $\{\alpha, \beta\}$, where α denotes the Dirichlet parameters of topic z and β is a matrix that contains value of the topic unigram $\beta_{w,z} = p(w|z)$. A topic mixture vector θ is drawn from the Dirichlet distribution with parameter α. The corresponding topic z is generated based on the multinomial distribution with parameter θ. Each word w_n is generated by the distribution $p(w_n|w_n, \beta)$. Finally, we obtain the marginal probability of document w by:

$$p(w|\alpha,\beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n=1}^{Z} p(z_n|\theta)p(w_n|z_n|\beta)d\theta \tag{4}$$

where N is size of document and Z is number of topic in document w. In [5], LDA probability of w_i was calculated by combining the topic probabilities with the topic- dependent unigram β:

$$p_{LDA}(w_i) = \sum_{z=1}^{Z} \beta_{w_i,z} \frac{\hat{\gamma}_z}{\sum_{j=1}^{Z} \hat{\gamma}_j} \tag{5}$$

where $\hat{\gamma}_z$ is variational parameter that approximated Bayes estimates for the LDA model via an alternating variational expectation maximization (EM) procedure [7].

2.2 Dirichlet Class Language Model

In [5], Dirichlet class language model (DCLM) is introduced, in which the class structure is estimated by Dirichlet densities from n-gram events. The class uncertainty is compensated by marginalizing the likelihood function over the Dirichlet priors. The latent variable in DCLM reflects the class of an n-gram event rather than the topic in LDA model, which is extracted from large-span documents. DCLM is considered as a kind of class-based LM. In contrast, the class label in a traditional class-based LM has been associated with an individual history word, and derived separately from the stage of model parameters. However, the class structure and the model parameters are consistently estimated under the same criterion in the proposed DCLM.

A linear discriminant function can be used to evaluate the contributions of the historical words to various classes. Without loss of generality, the $(n-1)V$ dimensional history vector h_{i-n+1}^{i-1} is projected into a c dimensional class space using a class-dependent linear discriminant function [8, 9]:

$$g_c\left(h_{i-n+1}^{i-1}\right) = a_c^T h_{i-n+1}^{i-1} \tag{6}$$

where a_c is a parameter. This function reflects the class posterior probability $p\left(c|h_{i-n+1}^{i-1}\right)$, which is essential for predicting the class information for unseen history. $A = [a_1, \dots, a_C]$ is a basis vector that in [9] is established to span the class space. DCLM constructs a Bayesian latent class LM by compensating for the uncertainty associated with the latent classes c or class mixtures θ. The class information c in DCLM is drawn from a history dependent Dirichlet prior $\theta = [\theta_1, \dots, \theta_C]^T \sim \mathrm{Dir}\left(g\left(h_{i-n+1}^{i-1}\right)\right)$.

The joint probability of word w_i, class c_i and class mixture vector θ conditioned on history h_{i-n+1}^{i-1} and DCLM parameters {A,β}, is computed by:

$$p\left(w_i, c_i, \theta|h_{i-n+1}^{i-1}, A, \beta\right) = p(w_i|c_i, \beta)p(c_i|\theta)p\left(\theta|h_{i-n+1}^{i-1}, A\right) \tag{7}$$

The parameters A, β were estimated using the variational Bayes expectation maximization (VB-EM) algorithm [8]. The n-gram probability obtained using DCLM is expressed in a form of marginal likelihood as:

$$p\left(w_i|h_{i-n+1}^{i-1}, A, \beta\right) = \sum_{c=1}^{C} \beta_{w_{ic}} \frac{g_c\left(h_{i-n+1}^{i-1}\right)}{\sum_{j=1}^{C} g_j\left(h_{i-n+1}^{i-1}\right)} \tag{8}$$

Comparing LDA with DCLM indicates that whereas LDA calculates the document probability, DCLM calculates the word probability given history words. DCLM performs the unsupervised learning of latent classes of n-gram events through the VB-EM procedure. DCLM differs from the class-based n-gram, in two ways: firstly, the classes of history words are determined according to the mutual information criterion, secondly, the corresponding classes represent the order of words.

2.3 Cache Dirichlet Class Language Model

In DCLM procedure, the class mixtures θ are drawn from history words w_{i-n+1}^{i-1} using the Dirichlet distribution with parameters $g(h_{i-n+1}^{i-1})$. The class probability $p(c_i|\theta)$ is calculated and the word w_i is predicted incorporating the multinomial parameters $\beta = \left\{\beta_{w_i c_i}\right\}$. However, the long-distance information beyond the n-gram window is not captured. In Cache DCLM (CDCLM) [5], in order to perform the large-span language modeling, the class information must be continuously updated. The class mixtures θ are not only generated from $n-1$ history words but also from the class information $c_1^{i-1} = (c_1, \dots, c_{i-1})$ of all preceding words w_1^{i-1}. For simplification, CDCLM only used of a best class sequence $\hat{c}_1^{i-1}$. In the new language model, the probability of an n-gram event is calculated as follows:

$$p\left(w_i | h_{i-n+1}^{i-1}, A, \beta, w_1^{i-1}\right) \cong \sum_{c=1}^{C} \beta_{ic} \frac{g_c\left(h_{i-n+1}^{i-1}\right) + \rho \sum_{t=1}^{i-1} \tau^{i-t-1} \delta(c, \hat{c}_t)}{\sum_{j=1}^{C} [g_j\left(h_{i-n+1}^{i-1}\right) + \rho \sum_{t=1}^{i-1} \tau^{i-t-1} \delta(j, \hat{c}_t)]} \quad (9)$$

The derivation of Equation (9) is similar to that of Equation (8). In Equation (9), a weighting factor $0 < \rho < 1$ is empirically introduced to balance the history words and the previous class sequence information. Additionally forgetting factor $0 < \tau < 1$ is applied to discount distant class information. The class associated with the farther word has a smaller impact on the word prediction. In other words, the class sequence is weighted. In the case of $\rho = 0$, CDCLM reduced to DCLM. If ρ is very large, CDCLM is comparable to a class based cache, which is different from the word-based cache in previous cache LMs [10].

3. Proposed Methods

3.1 Dirichlet Class Language Model Based on Part-of-Speech

As mentioned before, the previous adaptation methods extract latent semantic information such as topic dependency of words. In the languages with high morphology for example Farsi, using of the syntactic information such as part-of-speech (POS) along with semantic information can be useful [11].

Accordingly, we proposed an idea for using POS based on DCLM. DCLM acts as a Bayesian topic LM in which the prior density of topic variable is characterized by n-gram events. In our proposed model, we use POS information of previous words along with history words h_{i-n+1}^{i-1} for word clustering. The order of history factors is represented in f_{i-n+1}^{i-1}. Similar to DCLM, first, we declare a linear discriminant function. This function can be used to represent the cooperation of the history factors to different classes.

$$g_c(f_{i-n+1:i-1}^{1:2}) = a_c^T f_{i-n+1:i-1}^{1:2} \quad (10)$$

Linear function shows the class posterior probability $p(c|f_{i-n+1:i-1}^{1:2})$ where $f_i^{1:2}: \{f_i^1 = w_i, f_i^2 = p_i\}$. The first factor is the word and the second one is the POS of the word. a_c is the same as in DCLM.

Fig. 1 shows the graphical model of DCLM based on POS (DCLM_POS) for a text corpus that comprises of previous factor events. The class information c in DCLM _POS is drawn from the parameter θ. The joint probability of word w_i, class c_i and parameter θ, conditioned on history factors $f_{i-n+1:i-1}^{1:2}$ and DCLM_POS parameters {A,β}, is computed by:

$$p(w_i, c_i, \theta | f_{i-n+1:i-1}^{1:2}, A, \beta) = p(w_i|c_i, \beta) p(c_i|\theta) p(\theta|f_{i-n+1:i-1}^{1:2}, A) \quad (11)$$

where A, β parameters are the same as in DCLM.

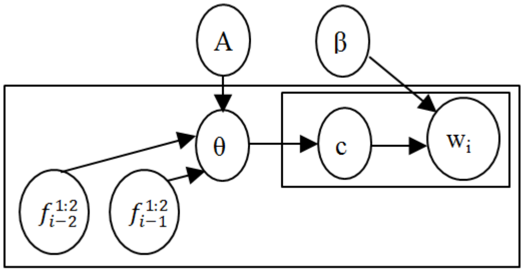

Fig. 1. Graphical representations for DCLM_POS

The n-gram probability based on previous factors is calculated by marginalizing the joint probability over the uncertainty of class mixture θ associated with different classes c_i:

$$p(w_i|f_{i-n+1:i-1}^{1:2}, A, \beta) = \sum_{c=1}^{C} \beta_{w_{ic}} \frac{g_c\left(f_{i-n+1:i-1}^{1:2}\right)}{\sum_{j=1}^{C} g_j\left(f_{i-n+1:i-1}^{1:2}\right)} \quad (12)$$

The DCLM_POS parameters are computed using the VB-EM procedure as in DCLM. After several VB-EM iterations, the DCLM_POS model inference converges.

Comparing LDA in (5) with DCLM_POS in (12) shows that whereas LDA calculates the word probability given word events in documents, DCLM_POS calculates the word probability from history factors.

3.2 Cache Dirichlet Class Language Model Based on Part-of-Speech

From the history factors $f_{i-n+1:i-1}^{1:2}$, the DCLM_POS first draws the class mixtures θ based on the Dirichlet distribution with parameters $g(f_{i-n+1:i-1}^{1:2})$. The word w_i in DCLM is predicted using the multinomial parameters $\beta = \{\beta_{w_i c_i}\}$. DCLM in DCLM_POS was substituted by CDCLM to perform the large-span language modeling and a new CDCLM_POS is developed. In this technique, the class information must be continuously updated. The class mixtures not only depend on $n-1$ history factors but also are influenced by the class information c_{i-2}^{i-1} of all preceding factors $f_{i-2:i-1}^{1:2}$. Fig. 2 shows the graphical model of CDCLM based on POS (CDCLM_POS) using two previous factors for model estimation.

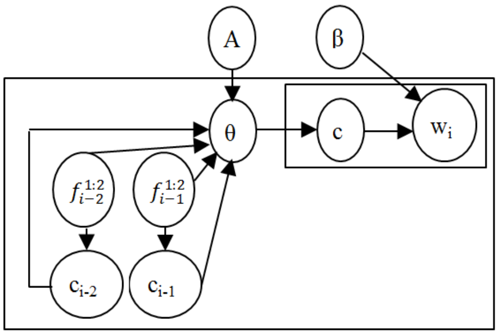

Fig. 2. Graphical representations for CDCLM_POS

As depicted in Fig. 2, to predict w_i, the class mixtures θ, are generated from the history factors $f_{i-2:i-1}^{1:2}$ and the class sequence c_{i-2}^{i-1}. Based on the CDCLM_POS, the probability of an n-gram event is calculated using:

$$p\left(w_i|f_{i-n+1:i-1}^{1:2}, A, \beta, w_1^{i-1}\right) \cong \sum_{c=1}^{C} \beta_{ic} \frac{g_c\left(f_{i-n+1:i-1}^{1:2}\right)+\rho\sum_{t=1}^{i-1}\tau^{i-t-1}\delta(c,\hat{c}_t)}{\sum_{j=1}^{C}[g_j\left(f_{i-n+1:i-1}^{1:2}\right)+\rho\sum_{t=1}^{i-1}\tau^{i-t-1}\delta(j,\hat{c}_t)]} \qquad (13)$$

where all of parameters are as in Equation (9).

4. Experiments

4.1 Dataset and Experimental Setup

The BijanKhan corpus [12] was utilized to evaluate the proposed methods in continuous speech recognition. The Farsdat training set was adopted to estimate the HMM parameters. The feature vector composed of 12 Mel-Frequency Cepstral Coefficients (MFCC) and one log energy and their first, second and third derivatives. Triphone models were built for 32 phones and each triphone model had three states with sixteen Gaussian mixtures in each state.

The HTK [13] was exploited for HMM training and lattice generation. The baseline LM was trained by SRILM [14] toolkit [1]. Kndiscount [15] method of smoothing methods used in the n-gram model. LDA toolkit[2] to train LDA model and DCLM toolkit[3] to train DCLM based models. In the experiments, number of topics Z and classes C were set to 100. The BijanKhan corpus with 10k documents, 70k distinct words and 40 POS was adopted to train the baseline LM comprised trigram model, FLM and proposed methods [16]. After removing the stop words, we used a lexicon with 45K frequent words for built the LDA model. In addition, the Farsdat corpus is 400 sentences. These corpuses were used to examine different models by perplexity and word error rate (WER). Firstly, we evaluated adapted LMs by perplexity criterion on the 10-fold procedure. Finally, in evaluation of speech recognition, we report WERs (%) of using different LMs.

Perplexity is the most common intrinsic evaluation metric for LM. A lower perplexity corresponds to less confusion in the prediction of language words. We apply 10-Fold mechanism to BijanKhan corpus for perplexity evaluation in all models. The general LM mixture approaches try to combine the topic or class model and traditional LM through some adaptation strategies [17]. Just in the way proposed by [18, 19], many methods are used to integrate the topic or class model with traditional LM which introduces a different type of information.

4.2 Experimental Results

In the experiments, we use linear interpolation for combining the trigram LM and FLM with adaptation methods. The interpolation weight between the basic LM and adaptation methods were determined from the 10-fold mechanism on perplexity metric. Fig. 3 shows the perplexities of the trigram LM, FLM and their linear combination with adaptation methods.

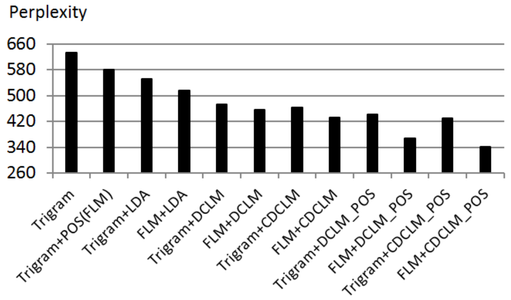

Fig. 3. Perplexities of models with linear interpolation

As Fig. 3 shows, the trigram LM and FLM had perplexity of 632 and 580 respectively. For FLM adaptation models with LDA, DCLM and CDCLM, the perplexities are about 516, 456 and 432 respectively. For FLM adaptation models with the proposed methods, DCLM_POS and CDCLM_POS, the perplexity is reduced to 369 and 342 respectively.

This experiment, represents that language model adaptation with techniques based on DCLM have significant improvement compared with adaptation based on LDA. Furthermore, word clustering has been improved using POS information of history words. In other words, using POS of previous words along with history words and class of history words for word clustering, improves the performance of trigram LM and FLM.

The evaluation of speech recognition was conducted using the Farsdat corpus. We reported the WERs (%) of various LMs. The HTK was used for acoustic model training using HMM and lattice generation. After that, the n-best list is created and then combined with estimated probability produced with LM using linear and log-linear combination. In this experiment, parameter n in n-best list is empirically set to five.

Fig. 4 shows the WERs of linear and log-linear combination of acoustic model with different LMs. This experiment represents that log-linear combination results in less WER than the linear combination.

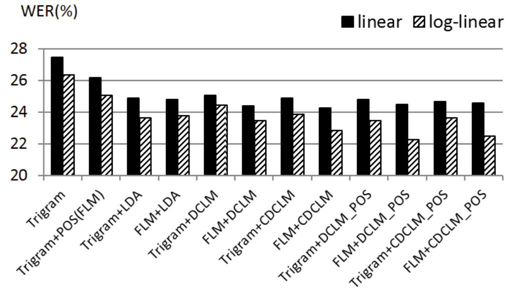

Fig. 4. WERs of combination LMs with acoustic model

[1] http://www.speech.sri.com/projects/srilm.
[2] http://www.cs.princeton.edu/~blei/lda-c.
[3] http://chien.csien.ncku.edu.tw/~dclm.

As depicted in Fig. 4, in the log-linear combination, trigram LM and FLM word error rate is about 26.4% and 25.1% respectively. The results of FLM adaptation are better than trigram LM. For FLM adaptation models with LDA, DCLM and CDCLM, word error rate is equal to 23.8%, 23.5% and 22.9% respectively. For FLM adaptation models with the proposed methods, DCLM_POS and CDCLM_POS, WERs is reduced to 22.3% and 22.5% respectively. Therefore, POS information can reduce WER in speech recognition system. Nevertheless, WER results confirm results of perplexity, but the trigram adaptation model with CDCLM_POS has less improvement than DCLM_POS.

The latent classes had been exploited by DCLM_POS methods. These classes were tagged here for ease of understanding. Fig. 5 displays some trigram events samples. Their coordinates in the class space that is spanned by the prior statistics of the previous factors $\{g_c(f_{i-n+1:i-1}^{1:2})\}$ on latent classes consist of POS {“Noun, Noun”, “Adjective, Noun”}.

The POS sequences {“Noun, Noun”, “Adjective, Noun”} are distributed in the top left and bottom right regions, respectively. The histories that are independent of these two classes are located in the bottom left region. The histories, “حَسن تصادف” and “حُسن تصادف” contained the same word sequences, but were located far apart in the class space.

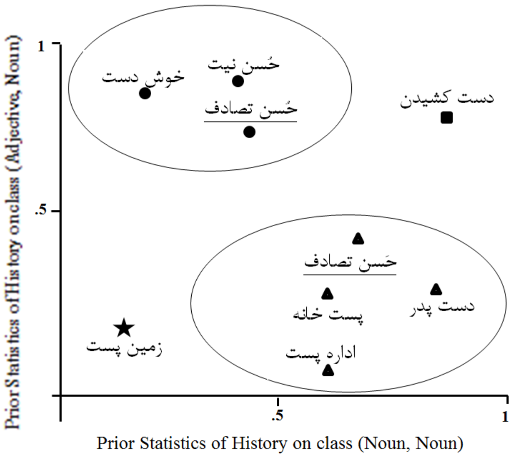

Fig. 5. Geometrical representations of latent class space constructed by the DCLM_POS

5. Conclusions

In summary, we have compared two approaches for using the part-of-speech (POS) information along with history words and class of history words. We use this information to cluster words and calculate the word probability. The first proposed technique is based on Dirichlet class language model (DCLM) using history words and POS of history words (DCLM_POS). The second is based on cache Dirichlet class language model (CDCLM) using history words, class of history words and POS of previous words (CDCLM_POS). Both methods are combined with trigram language model and factored language model in the form of linear interpolation. In this work, obtained language models are combined with acoustic model for speech recognition. In our experiments, the language model was build using the BijanKhan corpus, and the acoustic model was trained using Farsdat corpus. The lowest perplexity is achieved by linear combination of the factored language model and CDCLM_POS technique. The best word error rate is achieved using log-linear combination of the factored language model and DCLM_POS with acoustic model. As the future work, we will investigate the use of other linguistic features such as morphology. Using these features we hope to get improvements in the large variety of corpus. In addition, we study the discriminant functions in DCLM to reduce its computational complexity for online adaptation.

References

[1] L. R. Rabiner; R. W. Schafer, “Theory and application of digital speech processing,” Prentice-Hall, 2009.

[2] D. Jurafsky and J. H. Martin, “Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,” 2nd ed., Prentice Hall, 2008.

[3] K. Kirchhoff, J. Bilmes and K. Duh, “Factored Language Models Tutorial,” Department of Electrical Engineering University of Washington, 2008.

[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] J. T. Chien, and C. H. Chuen, “Dirichlet Class Language Models for Speech Recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, 2011.

[6] Y. C. Tam, T. Schultz, “Dynamic language model adaptation using variational Bayes inference,” *Proc. of EUROSPEECH*, pp. 5-8, 2005.

[7] S. Borman, “The expectation maximization algorithm a short tutorial,” Electronic document: www.seanborman.com/publications, 2004.

[8] J. T. Chien, and C. H. Chuen, “Latent Dirichlet Language Model for Speech Recognition,” *in Proc. IEEE Workshop Spoken Lang. Technol.*, pp. 201-204, 2008.

[9] J. T. Chien and C. H. Chuen, "Joint Acoustic and Language Modeling for Speech Recognition," *Speech Commun*., vol. 52, no. 3, pp. 223-235, 2010.

[10] P. R. Clarkson and A. J. Robinson, "Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache," *in Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process*., pp. 799-802, 1997.

[11] K. Kirchhoff, "Novel speech recognition models for Arabic," JHU, summer workshop final report, 2002.

[12] S. Tasharofi, F. Raja, F. Oroumchian and M. Rahgozar, "Evalution of Statistical Part of Speech Tagging of Persian Text," *International Symposium on Signal Processing and its Applications*, Sharjah, 2007.

[13] S. Young; G. Evermann; D. Kershaw; G. Moore; J. Odell; D. Ollason; D. Povey; V. Valtchev; P. Woodland, "The HTK book," Cambridge University Engineering Department, 2002.

[14] A. Stolcke, "SRILM-An extensible language modeling toolkit," *in Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, 2002.

[15] V. Siivola, "Language models for automatic speech recognition: construction and complexity control," dissertations in computer and information science for the degree of doctor of philosophy submitted to the Johns Hopkins University, 2007.

[16] A. Hatami, A. Akbari, B. Nasersharif, "Factored Language Model Adaptation Using Dirichlet Class Language Model for Speech Recognition," *Information and Knowledge Technology (IKT)*, PP. 438-442, 2013.

[17] Z. Lv, W. Liu and Z. Yang, "A New Language Model Adaptation Framework Using Modification of Structures of Background Corpus and Language Model," *Natural Language Processing and Knowledge Engineering*, pp. 1-4, 2009.

[18] j. R. Bellegarda, "Statistical Language Model adaptation: Review and Perspectives," *Speech Communication*, vol. 42, pp. 93-108, 2004.

[19] A. Gutkin, "Log-linear Interpolation of Language Models," Thesis, Department of Engineering, University of Cambridge, UK, 2000.

Ali Hatami received B.S. degree in computer engineering (Software) from the Islamic Azad University, Zanjan, Iran, in 2007. He received the M.S. degree in computer engineering (Artificial Intelligence) from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2012. He is currently Data expert in the Azmoon Keyfiat Company. His research interests include natural language processing, text processing, information retrieval and machine learning.

Ahmad Akbari received the B.S. degree in electronics engineering (1987) and M.S. degree in communication engineering (1989) from the Isfahan University of Technology, Isfahan, Iran. He received the Ph.D. degree in electrical engineering from the University of Rennes, Rennes, France, in 1995. He is Associate Professor in the Department of Computer Engineering, Iran University of Science and Technology. He is currently head of Research Center for Information Technology (RCIT).

Babak Nasersharif received the B.S. degree in hardware engineering from the AmirKabir University of Technology, Tehran, Iran, in 1997. He received M.S. and Ph.D. degree in computer engineering (Artificial Intelligence) from Iran University of Science and Technology, Tehran, Iran, in 2001 and 2007 respectively. He is currently Assistant Professor in the Electrical and Computer Engineering Department K.N. Toosi University of Technology.