UNSUPERVISED LANGUAGE MODEL ADAPTATION USING LATENT DIRICHLET ALLOCATION AND DYNAMIC MARGINALS

Md. Akmal Haidar and Douglas O'Shaughnessy

INRS-Energy, Materials and Telecommunications, University of Quebec
800 de la Gauchetiere Ouest, H5A 1K6, Montreal, Canada
haidar@.emt.inrs.ca , dougo@emt.inrs.ca

ABSTRACT

In this paper, we introduce an unsupervised language model adaptation approach using latent Dirichlet allocation (LDA) and dynamic marginals: locally estimated (smoothed) unigram probabilities from in-domain text data. In LDA analysis, topic clusters are formed by using a hard-clustering method assigning one topic to one document based on the maximum number of words chosen from a topic for that document. The n-grams of the topic generated by hard-clustering are used to compute the mixture weights of the component topic models. Instead of using all the words of the training vocabulary, selected words are used for LDA analysis, which are chosen by incorporating some information retrieval techniques. We adapted the LDA adapted topic model by minimizing the Kullback-Leibler (KL) divergence between the final adapted model and the LDA adapted topic model subject to a constraint that the marginalized unigram probability distribution of the final adapted model is equal to the dynamic marginals. We have compared our approach with the conventional adapted model obtained by minimizing the KL divergence between the background model and the adapted model using the above constraint. We have seen that our approach gives significant perplexity and word error rate (WER) reductions over the traditional approach.

1. INTRODUCTION

Language model (LM) adaptation plays an important role for many research areas such as speech recognition, machine translation, and information retrieval. 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].

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]. 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) [3], Probabilistic Latent Semantic Analysis (PLSA) [4], and LDA [5]. 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. However, the LDA model can be viewed as a mixture of unigram latent topic models. This LDA adapted unigram model is used for dynamic marginals to form an adapted model by minimizing the KL divergence between the background model and the adapted model, subject to a constraint that the marginalized unigram probability distribution of the adapted model is equal to the corresponding distribution obtained by the LDA adapted unigram model [6]. The idea of using dynamic marginals and the formation of the adapted model by minimizing the KL divergence between the background model and the adapted model is proposed in [7]. Here, the dynamic marginals are the unigram distribution obtained from in-domain text data. We used their approach in our work.

In this paper, we extend our previous work [8] to find an 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. For LDA analysis, we have chosen the words from the training vocabulary, incorporating some information retrieval techniques. We have removed the MIT stop words list [9] and the words that occur only once in the training set from the training vocabulary. Topic clusters are formed by using a hard-clustering method. 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 model [8]. The final adapted model is formed by minimizing the KL divergence between the final adapted model and the LDA adapted topic model, subject to a constraint that the marginalized unigram distribution of the final adapted model is equal to the unigram distribution estimated from some in-domain text data: called dynamic marginals [7]. We compared our approach with the traditional approach where the 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 have seen that our approach gives significant reductions in perplexity and word error rate.

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

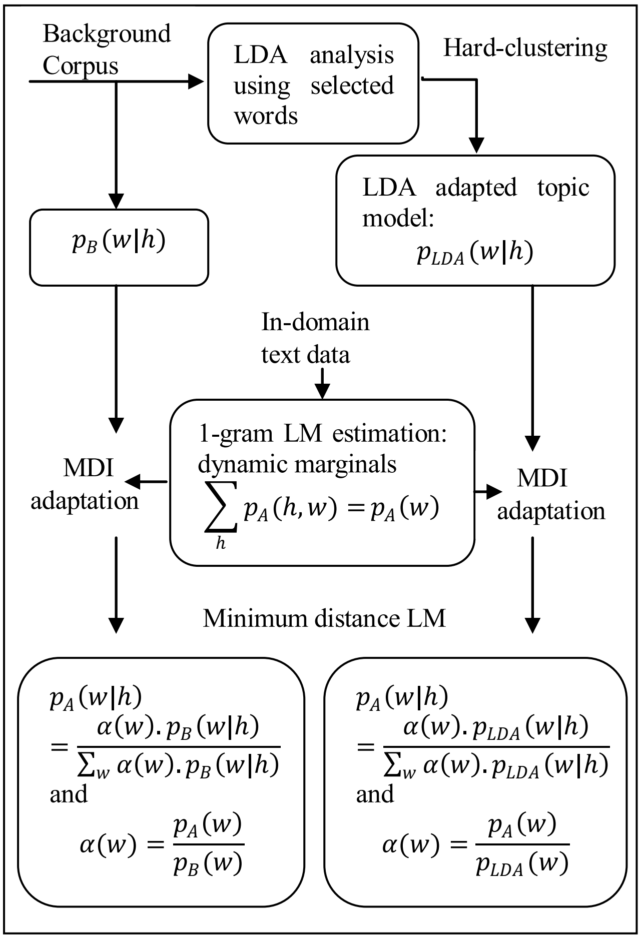
Background Corpus
LDA analysis using selected words
Hard-clustering
LDA adapted topic model: $p_{LDA}(w|h)$
$p_B(w|h)$
In-domain text data
1-gram LM estimation: dynamic marginals $\sum_h p_A(h,w) = p_A(w)$
MDI adaptation
MDI adaptation
Minimum distance LM
$p_A(w|h) = \frac{\alpha(w).p_B(w|h)}{\sum_w \alpha(w).p_B(w|h)}$ and $\alpha(w) = \frac{p_A(w)}{p_B(w)}$
$p_A(w|h) = \frac{\alpha(w).p_{LDA}(w|h)}{\sum_w \alpha(w).p_{LDA}(w|h)}$ and $\alpha(w) = \frac{p_A(w)}{p_{LDA}(w)}$

Figure 1: Unsupervised LM adaptation using LDA and MDI

2. RELATED WORK

To compensate for the weakness of *n*-gram models, which capture only short-range dependencies between words, many methods have been investigated. Cache-based language models are created based on the property that a word appearing earlier in a document is likely to occur again, which helps to increase the probability of the previously observed words in a document when predicting the next word [10]. This idea is used to increase the probability of unobserved but topically related words, for example trigger-based LM adaptation using a maximum entropy framework [11].

Recently, latent topic analysis has been introduced widely for language modeling. Topic clusters can be formed by using a hard-clustering method where a single topic is assigned to each document and used in LM adaptation [12]. Topics are extracted through a word-document co-occurrence matrix in LSA, which showed significant reductions in perplexity and WER in LM adaptation [3]. The probabilistic LSA (PLSA) is used to decompose documents into unigram topic models and combined with a generic tri-gram model to achieve perplexity reduction in LM adaptation [4]. PLSA cannot be used to model the unseen document as each document has its own set of topic mixture weights. So, for a large amount of documents, the parameter size increased significantly and the model is prone to overfitting. One of the most powerful probabilistic bag-of-words models is the LDA model, which imposes a Dirichlet distribution over topic mixture weights corresponding to the documents in the corpus. LDA provides a generative framework for explaining the probability of an unseen document. It overcomes the overfitting problem of PLSA by the limited number of model parameters that are dependent only on the number of topic mixtures and vocabulary size and achieves better results in perplexity reduction [5].

LDA has been successfully 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 [13]. The LDA model is used to extract topic clusters by using a hard-clustering method. The topic-specific tri-gram LM's are then combined with the generic tri-gram LM to obtain perplexity reduction [14]. The unigram count of the topic generated by hard clustering is used to compute the mixture weights of the topic models and has shown significant improvement in perplexity and WER reductions [15].

Many approaches have been proposed in the literature for language model adaptation using MDI. The idea is to minimize the KL divergence between the background model and the adapted model subject to a constraint that the marginalized unigram probability distribution of the adapted model is equal to the unigram distribution, which is estimated from some in-domain text data. The latter unigram distributions are called dynamic marginals [7]. Here, the author imposed an additional constraint to minimize the computational cost in computing the normalization term. The 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. The same technique is used in [6] and the LDA adapted unigram distribution is used as the dynamic marginal instead of using a locally estimated unigram distribution.

3. REVIEW OF LDA & TOPIC CLUSTERING

3.1. Latent dirichlet allocation

LDA is a popular probabilistic bag-of-words model [5]. 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,...,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 *θ(d).*

  - Choose a word *w* from the multinomial distribution *Φ(w | k ,β).*

where *α={α₁,...,α_K}* is used as the representation count for the *K* latent topics, θ indicates the relative importance of topics for a document and *Φ(w | k ,β)* represents the word probabilities conditioned on the topic with a Dirichlet prior and indicates the relative importance of particular words in a topic. *α* and *β* are Dirichlet priors that control the smoothing of the topic distribution and topic-word distribution, respectively [16].

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

$$p(d) = \int_{\theta} (\prod_{i=1}^{n} \sum_{k=1}^{K} \Phi(w_i \mid k, \beta) . \theta_k) f(\theta;\alpha) d\theta$$

3.2. Topic clustering

We have used the MATLAB topic modeling toolbox [17] 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_k 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_k (*k*=1,2…,*K*).

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

$$z_i^* = \underset{1 \le k \le K}{\operatorname{argmax}} DP(i,k) \qquad (1)$$

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.

4. LM ADAPTATION APPROACH

4.1. LDA adapted topic mixture model generation

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

$$P_{LDA}\ (w_k|h_k) = \sum_{i=1}^{K} \gamma_i\ p_{z_i}(w_k|h_k) \qquad (2)$$

where $p_{z_i}(w_k|h_k)$ is the i^{th} topic model and γ_i is the i^{th} mixture weight.

To find topic mixture weight γ_i , the *n*-gram count of the topics, generated by Equation (1) is used. Therefore,

$$\left.\begin{aligned} \gamma_k &= \sum_{j=1}^{n} P(z_k|w_{j-n},\ldots,w_{j-1}) P(w_{j-n},\ldots,w_{j-1}|d) \\ P(z_k|w_{j-n},\ldots,w_{j-1}) &= \frac{TF(w_{j-n},\ldots,w_{j-1},k)}{\sum_{p=1}^{K} TF(w_{j-n},\ldots,w_{j-1},p)} \\ P(w_{j-n},\ldots,w_{j-1}|d) &= \frac{freq(w_{j-n},\ldots,w_{j-1})}{Total\ counts\ of\ all\ n-grams} \end{aligned}\right\} \qquad (3)$$

where $TF(w_{j-n},\ldots,w_{j-1},k)$ represents the number of times the *n*-gram $w_{j-n},\ldots,w_{j-1}$ is drawn from topic z_k, 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.*

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

$$P(w_k|h_k) = \lambda * P_B(w_k|h_k) + (1-\lambda) * P_{LDA}(w_k|h_k) \qquad (4)$$

4.2. Adaptation using dynamic marginals

The adapted model using dynamic marginals [7] is obtained 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:

$$\sum_h p_A(h).p_A(w|h) = p_A(w) \qquad (5)$$

The constraint optimization problem has close connection to the maximum entropy approach [11], 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

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

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:

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

where β is a tuning factor between 0 and 1. In our experiments we used the value of β as 0.5 [6]. 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 model have standard back-off structure and the above constraint, so the adapted LM has the following recursive formula:

$$p_A(w|h) = \left\{ \begin{array}{ll} \frac{\alpha(w)}{z(h)}.p_{B/LDA}(w|h) & \text{if } (h,w)\text{exists} \\ b(h).p_A(w|\hat{h}) & \quad\text{otherwise} \end{array} \right\}$$

where

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

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(h)* is used to perform normalization similar to Equation (6), but the summation is taken only on the observed alternative words with the same word history *h* in the LM [6].

5. EXPERIMENTS AND RESULTS

5.1. Data and experimental setup

We evaluated the LM adaptation approach using the WSJ1 corpus transcription text data. We used all the training transcription text data for training and development and the 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. As 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 per document and in total 261 documents are created. The summary of training and testing datasets is given in Table 1.

Table 1: Summary of the Data Set

Corpus	Number of Words for Training	Number of Words for Testing	No. of words used for LM creation	No. of words used for LDA analysis
WSJ1 Transcription	1,317,793	Dev. Set:7235 Eval. Set:6708	20000	15282

We used the SRILM toolkit [18] and HTK toolkit [19] for our experiments. We trained LMs and used the *compute-best-mix* program from the SRILM toolkit to find the optimal weight λ 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 [20], 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 [20].

5.2. Perplexity reduction

We employed LDA on the WSJ1 training transcription text data to create 40 topic clusters. The bi-gram and 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 (3). The LDA adapted model is formed using Equation (2). Finally, the LDA adapted model is interpolated with the baseline model using Equation (4). Besides, we used dynamic marginals (unigram distribution of test sets) to adapt the background model and the LDA adapted topic model subject to the constraint in Equation (5). All the adapted models give significant perplexity reduction over background model. The results of the experiments are shown in Table 2. The language model in the second, third and fourth rows of the Table 2 shows significant reduction in perplexity of about 28.3% and 28.5%, 38.00% and 37.68%, and 54.46% and 54.47% for bigrams, and about 47.07% and 49.21%, 37.29% and 37.39%, and 65.69% and 67.36% for trigrams respectively over the baseline model for the WSJ1 development test set 1 and evaluation test set 1. We also note that for the development and evaluation test set 1, the proposed approach yields about 26.54% and 26.95%, and about 45.29% and 47.86% reductions in perplexity for bi-grams and trigrams respectively over the traditional approach of MDI adaptation. Moreover, the MDI adaptation of LDA adapted models outperforms the interpolated models of LDA adapted model and the background model.

Table 2: Perplexity results of the bi-gram model for the WSJ1 training transcription text.

Language Model	Perplexity Development test set 1	Perplexity Evaluation test set 1
Baseline	2-gram 608.08	637.25
	3-gram 771.14	849.41
Interpolated Model (*n*-gram weighting)	2-gram 435.97	455.60
	3-gram 408.10	431.35
Adapted Model obtained by using MDI adaptation of the background Model.	2-gram 377.00	397.10
	3-gram 483.56	531.79
Adapted Model obtained by using MDI adaptation of the LDA adapted model	2-gram 276.91	290.08
	3-gram 264.52	277.23

Table 3: WER results for the WSJ1 Development and Evaluation test set1

Language Model	WER(%) Development Test set 1	WER(%) Evaluation Test set 1
Baseline	24.97	26.43
Interpolated Model (bi-gram weighting)	22.75	23.52
Adapted Model obtained by using MDI adaptation of the background Model.	21.42	23.08
Adapted Model obtained by using MDI adaptation of the LDA adapted model.	19.98	21.08

5.3. Word error rate reduction

To evaluate the WER reduction using *HVite* in the HTK toolkit, we used only the bi-gram model. The results of the experiments are shown in Table 3. From the table, we note that all the adapted models outperform the baseline model. The language model in the second, third and fourth rows of the Table 3 gives significant WER reductions of about 8.89% and 11.01%, 14.21% and 12.67%, and 19.98% and 20.24% respectively over the baseline model for the WSJ1 development test set 1 and evaluation test set 1. We can also note that the proposed approach outperforms the traditional approach of MDI adaptation and the interpolated model.

6. CONCLUSION

We proposed an unsupervised language model adaptation approach using LDA and MDI. We computed the LDA adapted topic model by forming topic clusters from the background corpus using a hard-clustering method, and the *n*-gram weighting approach are used to compute the mixture weights of the component topic models. The adapted model using MDI is computed by minimizing the KL divergence between the adapted model and an LDA adapted topic model subject to a constraint that the marginalized unigram probability distribution of the adapted model is equal to some unigram distribution estimated from in-domain text data called dynamic marginals. We compared our approach with the traditional MDI adaptation of background model using the same constraint and have seen that our approach gives significant reductions in perplexity and WER. Furthermore, the MDI adaptation of LDA adapted model outperforms the interpolated model of the LDA adapted model and the background model in both perplexity and WER. However, the proposed MDI method needs extra computational cost for LDA analysis compared to the conventional MDI approach.

REFERENCES

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

[2] R. Kneser and V. Steinbiss, “On the Dynamic Adaptation of Stochastic Language Models”, in *Proc. of ICASSP*, vol. 2, pp. 586-589, 1993.

[3] 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.

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

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

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

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

[8] M. A. Haidar and D. O’Shaughnessy, “Unsupervised Language Model Adaptation Using *N*-gram weighting”, in *Proc. of CCECE,* 2011.

[9] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[10] 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.

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

[12] 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.

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

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

[15] 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.

[16] A. Sethy and B. Ramabhadran, “Bag-Of-Word Normalized N-gram Models”, in *Proc. of INTERSPEECH*, pp. 1594-1597, 2008.

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

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

[19] 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.

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