IMPROVING ARABIC BROADCAST TRANSCRIPTION USING AUTOMATIC TOPIC CLUSTERING

Stephen M. Chu and Lidia Mangu

IBM T. J. Watson Research Center
Yorktown Heights, New York, 10598, USA
{schu,mangu}@us.ibm.com

ABSTRACT

Latent Dirichlet Allocation (LDA) has been shown to be an effective model to augment n-gram language models in speech recognition applications. In this work, we aim to take advantage of the superior unsupervised learning ability of the framework, and use it to uncover topic structure embedded in the corpora in an entirely data-driven fashion. In addition, we describe a bi-level inference and classification method that allows topic clustering at the utterance level while preserving the document-level topic structures. We demonstrate the effectiveness of the proposed topic clustering pipeline in a state-of-the-art Arabic broadcast transcription system. Experiments show that optimizing LM in the LDA topic space leads to 5% reduction in language model perplexity. It is further shown that topic clustering and adaptation is able to attain 0.4% absolute word error rate reduction on the GALE Arabic task.

Index Terms— *Arabic ASR, topic clustering, language modeling*

1. INTRODUCTION

The *topic* is one of the fundamental semantic features in human language and speech. For language modeling, it provides an important domain for adaptation and optimization.

Considerable effort has been made to develop automatic clustering algorithms for partitioning training text into topic-specific sets. For instance, [1] uses a similarity measure based on inverse document frequencies (idf); [2] proposed latent semantic analysis (LSA); [3]-[4] use the mixture of unigrams model; and [5]-[6] applies the probabilistic latent semantic analysis (pLSA) [7]. Compared to these techniques, the more recent latent Dirichlet allocation (LDA) model [8] has been shown to be a superior alternative for document modeling and text classification, and has been considered in a wide range of applications from information retrieval to image processing.

A number of work that aim to bring LDA into the ASR domain have been reported. In [9]-[10], LDA is applied to construct a topic-dependent unigram LM that is subsequently used to supplement the regular *n*-gram LM through unigram rescaling. [11] uses an HMM-LDA model that combines the HMM and LDA models to separate syntactic words with local dependencies from topic-dependent content. And [12] extends the LDA framework to include history dependence so that it can be used to as a replacement of the conventional n-gram LM.

In this work, instead of directly incorporating LDA as a component of the language model, we try to take advantage of the superior unsupervised learning ability of the model, and use it to uncover the topic structure embedded in the training corpora in an entirely data-driven fashion. This process can be interpreted as constructing a topic subspace, and by projecting an unstructured collection of data onto this space, we will be able to separate and structure the data along dimensions with clear semantic interpretations. Topic adaption for a given task then becomes simply searching for an optimum in topic subspace, which gives the appropriate mixture of topics.

In addition, we describe a bi-level inference and classification method that allows topic clustering at the sentence level while preserving the document-level topic structures. The main goals of the experiment are to *a.* evaluate LDA as a tool for unsupervised topic discovery in the context of language modeling, and *b.* demonstrate the effectiveness of the proposed topic-adaptation pipeline in a state-of-the-art Arabic ASR system.

The rest of the paper is organized as follows. In section 2, we review the basic LDA formulation, section 3 covers the topic discovery experiments and the specifics of our implementation; section 4 discusses the topic clustering and LM rescoring results; and finally, conclusions and future work are given in section 5.

2. LDA FORMULATION

LDA can be viewed as a probabilistic graphic model that captures the following generative process for each of the M documents of a corpus:

choose $N \sim \text{poisson}(\xi)$
choose $\theta \sim \text{dir}(\alpha)$
for each of the N words w_n
  choose a topic $z_n \sim \text{multinomial}(\theta)$
  choose a word w_n from $p(w_n|z_n, \beta)$

where α is a k-dimensional vector that specifies topic priors through the Dirichlet random variable θ, $p(w_n|z_n, \beta)$ is a multinomial probability conditioned on the topic z_n; β is a $k \times V$ matrix, where $\beta_{ij} = p(w^j = 1| \ z^i = 1)$. The number of topics, k, and the size of the vocabulary V, are both assumed known and fixed.

Given α and β, the joint distribution of topic mixture θ, a sequence of N topics $\mathbf{z}$, and N words of a document, $\mathbf{w}$ is

$$p\left(\theta,\mathbf{z},\mathbf{w} \mid \alpha,\beta\right) = p\left(\theta \mid \alpha\right)\prod_{n=1}^{N} p\left(z_n \mid \theta\right) p\left(w_n \mid z_n,\beta\right). \qquad (1)$$

Note that for a given corpus, the model is fully parameterized by α and β. In a typical application, document $\mathbf{w}$ is observable, while θ and $\mathbf{z}$ are hidden variables. Thus, the basic inference problem of the model is to compute the posterior of the hidden variables given a document,

$$p\left(\theta,\mathbf{z} \mid \mathbf{w},\alpha,\beta\right) = p\left(\theta,\mathbf{z},\mathbf{w} \mid \alpha,\beta\right) \Big/ p\left(\mathbf{w} \mid \alpha,\beta\right). \qquad (2)$$

This posterior distribution is intractable for exact inference. However, a number of approximate inference algorithms are available,

including various sampling based methods and the variational method. In the subsequent experiments, we apply a parallelized version of the variational EM based on the recipe given in [13].

3. UNSUPERVISED TOPIC DISCOVERY

For the topic discovery and data clustering application concerned in this work, our goal is not to evaluate unseen text using models built on a separate training set. Rather, we treat LDA as an unsupervised learning method and use it to uncover semantic structure in the entire corpus. Thus, the *training*, or the parameter estimation process itself becomes equivalent to the topic discovery procedure. In fact, the variational parameters introduced for approximate inference lead to a straightforward solution for document classification.

3.1. Topic classification through variational EM

The variational interference algorithm for LDA essentially constructs a simpler graphical model characterized by

$$q\left(\theta, \mathbf{z} \mid \gamma, \phi\right) = q\left(\theta \mid \gamma\right)\prod_{n=1}^{N} q\left(z_n \mid \phi_n\right), \tag{3}$$

with free variational parameters γ and ϕ that can be optimized to give a lower bound of the log likelihood in the original model. Solution of the optimization problem leads to the following update equations,

$$\phi_{ni} \propto \beta_{iv} \exp\left(\Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right) \tag{4}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{5}$$

where Ψ is the first derivative of the $\log\Gamma$ function, computable via Taylor approximations.

Given a collection of documents, parameters α and β that maximize the marginal log likelihood of the data are found through the following iterative EM procedure:

E Step	for each document, find the optimizing values of the variational parameters $\{\gamma_d^*, \phi_d^*\}$
M Step	maximize the resulting lower bound on the log likelihood w.r.t. the model parameters α and β

Note that the optimal variational parameters $\{\gamma_d^*, \phi_d^*\}$ found using equations (5) and (6) are document-dependent. In particular, the Dirichlet parameter γ_d^* is a k-dimensional vector that can be interpreted as the weights of the k topics for a particular document d. Therefore, this intermediary variable from the variational EM procedure can be used directly to determine topic configuration of each document in the corpus. Given γ_d^*, the dominant topic τ_d of a document d is simply,

$$\tau_d = \arg\max_k \gamma_d^*. \tag{6}$$

Finally, the k topics uncovered through the unsupervised learning process are given by the $k \times V$ matrix β, which specifies the unigram probabilities given topics.

3.2. Bi-level inference

The topic classification method given in equation (6) assigns a single topic to each document. Note that this is equivalent of a classifier that chooses the nearest apex on the $(k-1)$ -dimensional topic-simplex, which is a reasonable approximation when one topic is clearly dominant in the document. However, when multiple topics exhibit similar weights in γ_d^*, the classification will be much less reliable. In general, the mixture of topics in a document becomes separable when we consider each sentence separately. It is therefore more desirable to perform topic classification at the sentence-level.

However, simply treating the sentences as documents and applying an additional pass of inference using the LDA model is not optimal, as this jettisons the correct document/topic structure embedded in the training data. Indeed, we wish to preserve the original document-level topic mixture while computing the sentence-level statistics, and make classification decisions based on both levels.

Given a sentence $\mathbf{s} = \{w_1, w_2, \ldots w_n\}$ in document d, for each word w_i and each topic k, we compute,

$$g_k(w_i) \propto \exp\left\{\log \beta_{k,w_i} + \mathbb{E}[\log \theta_{d,k}]\right\}. \tag{7}$$

The first term in the summation is the log probability of word w_i given topic k; and the second term is the log probability of topic k in document d, which can be readily computed using the variational parameter γ_d^*,

$$\mathbb{E}[\log \theta_{d,k}] = f(\gamma_d^*). \tag{8}$$

The topic $\tau_{\mathbf{s}}$ of the sentence $\mathbf{s}$ is then

$$\tau_{\mathbf{s}} = \arg\max_k \prod_{i=1}^{n} g_k(w_i). \tag{9}$$

3.3. Topics discovery on Arabic Gigaword corpus

The described topic discovery and clustering methods were evaluated on the Arabic Gigaword corpus [14], a large archive of Arabic newswire data released by LDC.

An important assumption made in the LDA framework is that the Dirichlet random variable θ, which specifies the topic mixture, is a document-level parameter, thus it should remain constant for a given document. Therefore, the choice of document boundaries within a corpus will have considerable influence on both the stability and the quality of the learning outcome.

Most existing applications of LDA rely on document boundaries defined by the corpus. This approach is desirable when these boundaries are available and have a clear semantic justification, e.g., a collection of papers from scientific journals. However, for speech applications, the concept of document is often less well-defined, or lacks consistency in terms of length and granularity across a large collection from multiple sources. Therefore, it is useful to have a general mechanism to automatically process text into segments with controlled consistency.

The definition of utterance in human speech has arguably a much higher consistency than document across different tasks, applications, or even languages. One possibility then is to use each utterance as a document in LDA learning. However, this approach defeats one of the main merits of the LDA model that it allows a mixture of topics for each document. Practically, on the other hand, having documents with shorter lengths can reduce overall computational load in the variational EM procedure. It can be shown that the M-step is computationally inexpensive compared to the E-step, which is an iterative procedure itself, and has a total complexity on the order of MN_{max}^2K in each EM integration, where M is the number of documents and N_{max} is the maximum document length in the corpus. Because the complexity is quadratic in N_{max} and linear in M, it is computationally beneficial to divide the corpus

into higher number of documents with shorter length. In addition, it is desirable to have documents with consistent lengths so that N_{max} is minimized.

The Gigaword corpus indeed provides both document and utterance-level annotations. At the document level, there are three types of definitions. Instead of relaying on these markers as document boundaries, we treat the entire corpus as a continuous sequence of utterances, and use a moving window with a fixed size to partition the corpus into a set of *pseudo-documents*. Here we make the underlying assumption that the order of utterances in the corpus is not random.

Compared to English, Arabic has a substantially larger vocabulary. For instance, the Gigaword set used here has a list of more than 1.17M unique words. Left untreated, this would pose a hefty challenge in both computational load and model size, as the complexity of the variational EM algorithm grows linearly with the size of vocabulary. In our experiments, we constructed various much smaller vocabularies with sizes of 200k, 100k, 50k, and 20k by removing infrequent words from the list. We found that LDA is able to give robust topic estimation in Arabic text even with the drastic 20-to-1 vocabulary reduction as in the 50k case.

LDA is based on the bag-of-words assumption, which neglects the order of words in a document. This allows us to further remove the most frequent words directly from the data to reduce feature size without hampering the topic modeling performance.

A partition from the Gigaword corpus containing 223M words and 9.29M utterances is used in the topic discovery experiment. In the reported system, we apply a 100-utterance moving window with no overlap to partition the corpus into 93K pseudo-documents. Using the 50k vocabulary reduces the number of words to 195M; removing the 16 most frequent words further reduced the data size by 35.3M words. Variational EM parameter estimation with $k = 20$ is carried out on the resulting set until a target ratio of likelihood change between iterations is reached.

As discussed in the previous section, the LDA parameter β is a $k \times V$ matrix that specifies the unigram probabilities given topics. Therefore, looking into β will give a qualitative assessment of the topic discovery outcome. Table 1 shows the top words sorted by the corresponding conditional unigram for three of the topics in the final model. The primary English translation for each word is also included. Indeed, each of the three word lists exhibits a distinct semantic focus, which matches well with our topic discovery objective.

Table 1 Words with highest unigram probabilities in final β.

Topic 1		Topic 2		Topic 3	
المئة	percent	امام	front	اسرائيل	Israel
الدول	States	المباراة	match	الفلسطينية	Palestinian
النفط	oil	الاتحاد	Union	الفلسطيني	Palestinian
الاقتصادي	economic	الاول	first	الاسرائيلي	Israel
الاقتصاد	economy	كأس	Cup	الاسرائيلية	Israel
الاقتصادية	economic	بطولة	starring	الفلسطينيين	Palestinians
العام	General	القدم	football	غزة	Gaza
دولار	dollars	الثاني	II	حماس	Hamas
دول	States	الاولى	first	القدس	Jerusalem
عام	year	قبل	before	عرفات	Arafat
اسعار	prices	صفر	zero	السلطة	power
المالية	financial	مباراة	match	الضفة	West
السوق	market	المنتخب	team	شارون	Sharon
خلال	during the	كرة	Football	الحكومة	government
العالم	world	الاهلي	Al-Ahli	رئيس	President
بليون	one billion	العالم	world	السلام	peace

4. LM OPTIMIZATION EXPERIMENTS

The ultimate goal for unsupervised topic discovery here is to uncover semantic structure in data so that it can be used to improve language modeling performance in terms of perplexity and speech recognition performance.

4.1. LM optimization

The topic discovery process can be interpreted as constructing a topic subspace, and by projecting an unstructured collection of data onto this space, we are able to separate and structure the data along dimensions with clear semantic interpretations. Topic adaption for a given task then becomes simply searching for an optimum in topic subspace.

A straightforward recipe is to first build *k* topic-specific component LMs using the documents in the corresponding topic cluster as training data; then find the optimal set of interpolation weights for a specific task. Thus, instead of making the distinction of *on-topic* or *off-topic*, topic adaptation is expressed through optimizing a vector in the *topic space*. A deciding factor for the usefulness of the topic space is how *expressive* it is, which is determined by how well the component topic LMs are separated. This can be indirectly measured by the dynamic range observed in interpolation weights after optimization. A larger range usually indicates a higher degree of expressiveness.

For LM training, a 795K vocabulary is used. Component topic LMs are built on the 20 data clusters obtained using the described document classification method. For comparison, two more sets of partitions are produced from the same corpus. The first is generated by dividing the data sequentially into 20 subsets, and the second by dividing the data randomly.

The *Model-M*, a class-based exponential model [15], is used for all LM builds. We build both a 3-gram and a 4-gram version of each model. Otherwise all LMs share the same configurations. The 20 LMs in each case are linearly interpolated with the interpolation weights chosen to optimize perplexity on a held-out set.

The perplexity results are summarized in Table 2. In both the 3-gram and the 4-gram case, the interpolated LMs built on random partitions (*rnd*) and the sequential partitions (*seq*) show comparable perplexities, while the topic-space LM (*topic*) achieves consistent perplexity reductions.

Table 2 Perplexity Results

	3-gram	*4-gram*
rnd	1725.89	1694.36
seq	1723.72	1687.74
topic	1665.47	1609.49

For instance, compared to *rnd*, the topic-space LM reduces the perplexity by more than 5% in the 4-gram case. The corresponding interpolation weights are shown in Fig. 1. As the graph indicates, the interpolation weights for the LMs built on randomly partitioned training sets remain largely uniform; whereas significant degree of dynamics is demonstrated in the topic LM weights, an indication that the topic space provides an effective basis for LM optimization.

4.2. LM rescoring experiments

LM rescoring experiments were carried out on a state-of-the-art Arabic broadcast transcription system developed for the DARPA

GALE evaluations. For acoustic modeling, a speaker-adapted, Buckwalter vowelized *subspace Gaussian mixture model* (SGMM) built through both feature-space and model-space discriminative

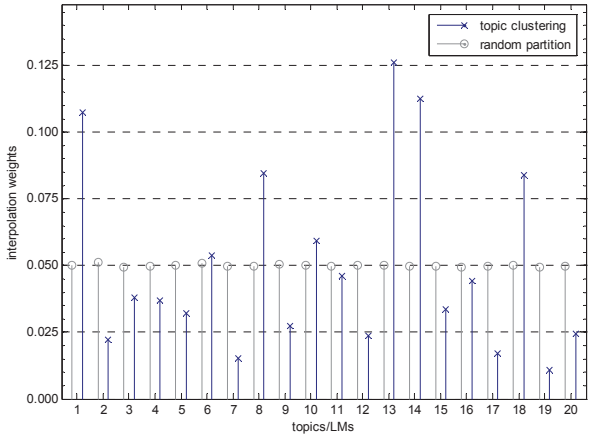

Fig. 1 The interpolation weights for the LMs built on randomly partitioned training sets remain largely uniform; whereas significant degree of dynamics is demonstrated in the topic LM weights, an indication that the proposed topic clustering provides an effective basis for LM optimization.

training on a acoustic training set with approximately 1800 hours of transcribed Arabic broadcasts data is employed. The SGMM model has 6K states and 150M Gaussians represented by an efficient subspace tying scheme. Details of the system are given in [16]. The LM used for constructing the static, finite-state decoding graph for lattice generation is trained on a large data set with 1.6 billion words, and subsequently pruned to 7M n-grams.

The performance of the interpolated LMs are evaluated on three shared test sets from the GALE program, referred to as *dev'07*, *dev'08*, and *eval'08* here. Rescoring results using the 4-gram version of the LMs are shown in Table 3. The results show that LM optimization in the topic space leads to a 0.2% absolute word error rate reduction on all three test sets, compared to LM built on random partitions of the same training data.

Table 3 Rescoring results (% WER)

	dev'07	*dev'08*	*eval'08*
rnd	9.1	10.6	9.7
topic	8.9	10.4	9.5
topic.bi-level	8.8	10.4	9.3

The proposed bi-level inference method is applied to cluster the data at the sentence level and create a new 20-subset partition. Compared to document-level topic clustering, 31.85% of the sentences and 29.09% of the words in the entire corpus are given different topic classifications. Again, 20 component LMs are built with the same basic configurations on the new partition, and corresponding interpolation weights are computed. The rescoring results (*topic.bi-level*) show that on two of the three test sets, bi-level inference is able to achieve further performance gains on top of the document-level clustering approach. For instance, on *eval'08*, bi-level inference gives an additional 0.2% absolute WER improvement, which brings the total error reduction to 0.4%.

5. CONCLUSIONS AND ACKNOWLEDGMENTS

The LDA framework is considered in the context of unsupervised topic discovery for language model optimization. We demonstrate an effective bi-level inference recipe for topic classification and validate the clustering/optimization pipeline on a state-of-the-art Arabic ASR system. Future work includes incorporating non-linear LM combination methods and optimizing interpolation weights using topic classifications on the test data. This work was supported in part by DARPA under grant HR0011-06-2-0001[§].

REFERENCES

[1] R. Iyer and M. Ostendorf, "Modeling long distance dependency in language: topic mixtures vs. dynamic cache models," in *Proc. ICSLP*, 1996.

[2] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling, " *Proceedings of IEEE*, vol. 88, no. 8, 2000.

[3] S. Martin, J. Liermann, and H. Ney, "Adaptive topic-dependent language modeling using word-based varigrams," in *Proc. Eurospeech*, 1997.

[4] R. Kneser and J. Peters, "Semantic clustering for adaptive language modeling," in *Proc. ICASSP*, 1997.

[5] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. Eurospeech*, 1999.

[6] D. Mrva and P. C. Woodland, "A pLSA-based language model for conversational telephone speech," in *Proc. ICSLP*, 2004.

[7] T. Hofmann, "Probabilistic Latent Semantic Analysis," *Uncertainty in Artificial Intelligence*, 1999.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *The Journal of Machine Learning Research*, 2003.

[9] Y. C. Tam and T. Shultz, "Correlated latent semantic model for unsupervised lm adaptation," in *Proc. ICASSP*, 2007.

[10] D. Mrva and P. C. Woodland, "Unsupervised language model adaptation for mandarin broadcast conversation transcription," in Proc. ICSLP, 2006.

[11] B. J. Hsu and J. Glass, "Style & topic language model adaptation using HMM-LDA," in *Proc. EMNLP*, 2006.

[12] J.-T. Chien and C.-H. Chueh, "Latent Dirichlet language model for speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 3, 2011.

[13] R. Nallapati, W. Cohen, and J. Lafferty, "Parallelized variational EM for latent Dirichlet allocation: an experimental evaluation of speed and scalability," in *Proc. ICDM workshop on high performance data mining*, 2007.

[14] R. Parker, et al., *Arabic Gigaword Fourth Edition*, Linguistic Data Consortium, Philadelphia, 2009.

[15] S. F. Chen, "Shrinking exponential language models," in *Proc. NAACL-HLT*, 2009.

[16] G. Saon, H. Soltau, U. Chaudhari, S. Chu, B. Kingsbury, H.-K. Kuo, L. Mangu, and D. Povey, "The IBM 2008 GALE Arabic speech transcription system," in *Proc. ICASSP*, 2010.

[§] Approved for Public Release, Distribution Unlimited. The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.