Modeling Topic and Role Information in Meetings using the Hierarchical Dirichlet Process

Modeling Topic and Role Information in Meetings using Hierarchical Dirichlet Process

Songfang Huang and Steve Renals

The Centre for Speech Technology Research
University of Edinburgh, Edinburgh, EH8 9LW, UK
{s.f.huang, s.renals}@ed.ac.uk

Abstract. In this paper, we address the modeling of topic and role information in multiparty meetings, via a nonparametric Bayesian model called hierarchical Dirichlet process, which provides a powerful solution to topic modeling and a flexible framework to incorporate other multimodal cues, i.e., the role information. We present our modeling framework for topic and role on the AMI Meeting Corpus, and illustrate the effectiveness of our approach in context of adapting a baseline language model in a large-vocabulary automatic speech recognition system for meeting, where it shows significant improvements in term of both perplexity and word error rate.

1 Introduction

A language model (LM) aims to provide a predictive probability distribution for the next word based on a history of previously observed words. The dominant LM for many state-of-the-art automatic speech recognition (ASR) systems nowadays is the conventional n-gram model, which approximates the history as the immediately preceding $n-1$ words. Although the n-gram model has been demonstrated to be simple but efficient approach to language modeling, the struggle to improve its performance always continues. Broadly speaking, there are two directions for those attempts. One tries to extend the $n-1$ word history to a richer context, while still remaining computationally feasible. Information used to extend the history includes morphological information in factored LMs [1], syntactic knowledge using structured LMs [2], and semantic knowledge such as topic information using Bayesian models [3]. Other attempts focus on different interpretations from the maximum likelihood estimated n-gram, such as neural networks [4], latent variable models [5], and a Bayesian framework [6, 7].

In this paper, we look at an improved LM for ASR in meetings by the inclusion of richer knowledge from multiparty meetings into a conventional n-gram model. The meeting corpus we consider here is the AMI Meeting Corpus[1] [8], which consists of 100 hours of multimodal meeting recordings with comprehensive annotations at a number of different levels. About 70% of the corpus was

[1] http://corpus.amiproject.org

elicited using a design scenario, in which the participants play the roles of employees, i.e., project manager (PM), marketing expert (ME), user interface designer (UI), and industrial designer (ID), in an electronics company that decides to develop a new type of television remote control. Our work in this paper is motivated by the fact that the AMI Meeting Corpus has a wealth of multimodal information such as audio, video, lexical, and other high-level knowledge. From the viewpoint of language modeling, the question for us is whether there are some multimodal cues besides lexical information would be helpful for improving a n-gram LM. If so, then what are those cues, and how could we incorporate them into a n-gram? To address this question, we have a focus on the modeling of topic and role information using a hierarchical Dirichlet process [9].

We consider an augmented n-gram model for ASR, with its context enriched by the inclusion of two multimodal cues from meeting: the *topic* and the speaker *role*. Unlike the role, which could be seen as deterministic information available in the corpus, the topic here refers to the semantic context, which is typically extracted by an unsupervised approach. One popular topic model is latent Dirichlet allocation (LDA) [10], which has proven to be a successful approach to automatically find the latent topics based on the co-occurrences of words in a 'document'. However, there are two difficulties posing with the application of LDA to language modeling. First, it is important to define a suitable document for LDAs to be used with LMs, because the data for language modeling tasks normally consist of sequences of short sentences, which do not fall in well-defined documents. Second, it is not easy to decide the number of topics, which is required to be set in advance for LDA.

More recently, a nonparametric generalization of LDA called the hierarchical Dirichlet process (HDP) [9] has been proposed. The HDP extends the standard LDA in two folds. First, the use of a Dirichlet process as a prior for the topic distribution, rather than the Dirichlet distribution in LDA, enables the HDP to determine the number of topics required. Second, the hierarchical (tree) structure enables the HDP to share mixture components (topics) between groups of data. In this paper we exploit the HDP as our modeling approach for automatic topic learning. Moreover, we also find it easier for us to incorporate roles together with topics by expressing them as an additional level of variables into the HDP hierarchy.

Some previous work has been done in the area of combining n-gram models and topic models such as LDA and probabilistic latent semantic analysis (pLSA) for ASR on different data, for example, broadcast news [11, 12], and lecture recordings [13]. The new ideas we exploit in our work cover the following aspects. Firstly, we use the nonparametric HDP for topic modeling to adapt n-gram LMs. Secondly, we consider sequential topic modeling, and define documents for the HDP by placing a moving window over the sequences of short sentences. Thirdly, we incorporate the role information with topic models in a hierarchical Bayesian framework. In the rest of this paper, we will review topic models, and introduce our framework for modeling topic and role information using HDP, followed by a set of perplexity and WER experiments.

2 Probabilistic Topic Model

Topic models, which recently received growing interest in the machine learning community, have been proposed for document modeling to find a latent representation (topic) connecting documents and words. In a topic model, words in a document exchangeably co-occur with each other according to their semantics meanings, following the "bag-of-words" assumption.

Suppose there are D documents in the corpus, and W words in the vocabulary. Each document $d = 1, \ldots, D$ in the corpus is represented as a mixture over latent topics (let $\boldsymbol{\theta}_d$ be the mixing proportions over topics), and each topic $k = 1, \ldots, K$ in turn is a multinomial distribution over words in the vocabulary (let $\boldsymbol{\phi}_k$ be the vector of probabilities for words in topic k).

In this section, we review two "bag-of-word" models, LDA and HDP, based on [9, 14, 15].

2.1 Latent Dirichlet Allocation

Latent Dirichlet allocation [10] is a three-level hierarchical Bayesian model, which pioneered the use of Dirichlet distribution for latent topics. That is, the topic mixture weights $\boldsymbol{\theta}_d$ for the dth document are drawn from a prior Dirichlet distribution with parameters $\alpha, \boldsymbol{\pi}$:

$$P(\boldsymbol{\theta}_d|\alpha\boldsymbol{\pi}) = \frac{\Gamma(\sum_{i=1}^{K}\alpha\pi_i)}{\prod_{i=1}^{K}\Gamma(\alpha\pi_i)}\theta_1^{\alpha\pi_1-1}\ldots\theta_K^{\alpha\pi_K-1} \tag{1}$$

where K is the predefined number of topics in LDA, Γ is the Gamma function, $\alpha\boldsymbol{\pi} = \{\alpha\pi_1, \ldots, \alpha\pi_K\}$ represents the prior observation counts of the K latent topics with $\alpha\pi_i > 0$, $\boldsymbol{\pi}$ is the corpus-wide distribution over topics, and α is called the concentration parameter controlling the amount of variability from $\boldsymbol{\theta}_d$ to their prior mean $\boldsymbol{\pi}$.

Similarly, Dirichlet priors are placed over the parameters $\boldsymbol{\phi}_k$ with the parameters $\beta\boldsymbol{\tau}$. We have,

$$\boldsymbol{\theta}_d|\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha\boldsymbol{\pi}) \qquad \boldsymbol{\phi}_k|\boldsymbol{\tau} \sim \mathrm{Dir}(\beta\boldsymbol{\tau}) \tag{2}$$

Fig. 1.(A) depicts the graphical model representation for LDA. The generative process for words in each document are as follows: first draw a topic k with probability θ_{dk}, then draw a word w with probability ϕ_{kw}. Let w_{id} be the ith word token in document d, and z_{id} the corresponding drawn topic, then,

$$z_{id}|\boldsymbol{\theta}_d \sim \mathrm{Mult}(\boldsymbol{\theta}_d) \qquad w_{id}|z_{id}, \boldsymbol{\phi}_{z_{id}} \sim \mathrm{Mult}(\boldsymbol{\phi}_{z_{id}}) \tag{3}$$

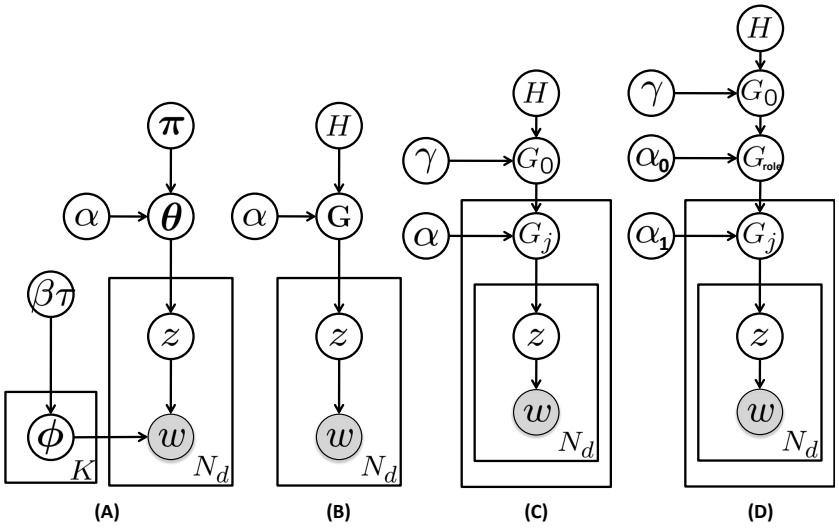

Fig. 1. Graphical model depictions for (A) latent Dirichlet allocation (finite mixture model), (B) Dirichlet process mixture model (infinite mixture model), (C) 2-level hierarchical Dirichlet process model, and (D) the role-HDP where G_{role} denotes the DP for one of the four roles (PM, ME, UI, and ID) in the AMI Meeting Corpus. Each node in the graph represents a random variable, where shading denotes an observed variable. Arrows denote dependencies among variables. Rectangles represent plates, or repeated sub-structures in the model.

2.2 Hierarchical Dirichlet Process

LDA pioneered the use of Dirichlet distributed latent variables to represent shades of memberships to different cluster or topics, while the HDP pioneered the use of nonparametric models to sidestep the need for model selection [15]. Two extensions were made by the HDP: firstly Dirichlet distributions in LDA are replaced by Dirichlet processes in the HDP as priors for topic proportions, and secondly priors are arranged into hierarchical tree structure.

Dirichlet Process. The Dirichlet process (DP) is a stochastic process first formalised in [16] for general Bayesian modeling, which has become an important prior used for nonparametric models. Nonparametric models have their number of model parameters growing with the amount of data, which helps to alleviate over- or under-fitting problems, and provide an alternative approach to parametric model selection or averaging.

A random distribution G over a space Θ is called a Dirichlet process distributed with base distribution H and concentration parameter α, if

$$(G(A_1), \ldots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)) \tag{4}$$

for every finite measurable partition $A_1, \ldots, A_r$ of Θ. We write this as $G \sim \mathrm{DP}(\alpha, H)$. The parameter H, a measure over Θ, is intuitively the mean of the DP. The parameter α, on the other hand, can be regarded as an inverse variance of its mass around the mean H, with larger values of α for smaller variances. More importantly in infinite mixture models, α controls the expected number of mixture components in a direct manner, with larger α implying a larger number of mixture components a priori.

Draws from a DP are composed as a weighted sum of point masses located at the previous draws $\theta_1, \ldots, \theta_n$. This leads to a constructive definition of the DP called the stick-breaking construction [17]:

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \qquad \pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_k) \qquad \theta_k^* \sim H \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \tag{5}$$

Then $G \sim \mathrm{DP}(\alpha, H)$. θ_k^* is the unique values among $\theta_1, \ldots, \theta_n$. The construction of $\boldsymbol{\pi}$ can be understood as follows [14]. Starting with a stick of length 1, first break it at β_1, assign π_1 to be the length of stick just broken off. Then recursively break the other portion to obtain π_2, π_3 and so forth. The stick-breaking distribution over $\boldsymbol{\pi}$ is sometimes written as $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$[2], and satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability one. This definition is important for the inference for the DP.

Recall in Equation 2 for LDA, a finite-dimensional Dirichlet distribution (i.e., of which $\boldsymbol{\pi}$ is a K-dimensional vector) is used as prior for distribution of topic proportions. LDA, in this sense, is a finite mixture model. If we use a DP instead as prior for mixing topic proportions, that is, $\boldsymbol{\theta}_d \sim \mathrm{DP}(\alpha, H)$ where $\boldsymbol{\phi}_k | H \sim \mathrm{DP}(\beta\boldsymbol{\tau})$, then the stick-breaking construction for $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ will produce a countably infinite dimensional vector $\boldsymbol{\pi}$. In this way, the number of topics in this DP-enhanced LDA model is potentially infinite, the number of topics increasing with the available data.

This model, as shown in Fig. 1.(B), is called the Dirichlet process mixture models (also known as infinite mixture model).

Hierarchical Framework. Besides the nonparametric extension of LDA from Dirichlet distribution to Dirichlet process, [9] further extends the Dirichlet process mixture model from a flat structure to a hierarchical structure, called a hierarchical Dirichlet process mixture model. This extended model use the hierarchical Dirichlet process as priors. Similar to the DP, the HDP is a prior for nonparametric Bayesian modeling. The difference is that in HDP, it is assumed that there are groups of data, and infinite mixture components are shared among groups.

Considering a simple 2-level HDP as an example, as shown is Fig. 1.(C)., HDP defines s set of random probability measure G_j, one for each group of data, and a global random probability measure G_0. The global measure G_0 is distributed as a DP with concentration parameter γ and base probability measure H, and

[2] GEM stands for Griffiths, Engen, and McCloskey.

the random measure G_j, assuming conditionally independent given G_0, are in turn distributed as a DP with concentration parameter α and base probability measure G_0:

$$G_0|\gamma, H \sim \mathrm{DP}(\gamma, H) \qquad G_j|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0) \tag{6}$$

This results in a hierarchy of DPs, in which dependencies are specified among s set of DPs by arranging them into a tree structure. Although this is a 2-level example, the HDP can readily be extended to as many levels as required.

A HDP-enhanced LDA model, therefore, will have a potentially infinite number of topics, and these topics will be shared among groups of data. If a HDP is used as prior for topic modeling, then the baseline distribution H provides the prior distribution for words in the vocabulary, i.e., $\boldsymbol{\phi}_k|H \sim \mathrm{DP}(\beta\boldsymbol{\tau})$. The distribution G_0 varies around the prior H with the amount of variability controlled by γ, i.e., $G_0 \sim \mathrm{DP}(\gamma, \mathrm{Dir}(\beta\boldsymbol{\tau})$. The actually distribution G_d for dth group of data (words in dth document in topic models) deviates from G_0, with the amount of variability controlled by α, i.e., $G_d \sim \mathrm{DP}(\alpha, G_0)$. Together with (3), this completes the definition of a HDP-enhanced LDA topic model.

3 Modeling Topic and Role using HDP

We emphasize in this section three key questions concerning with the modeling of topic and role using HDP. First, how to define a document in a multiparty meeting? Second, how to introduce role into the HDP framework? Third, how to use the local estimates from HDP to adapt a baseline n-gram LM for a ASR system?

Define a Document. The target application of the HDP in this paper is to adapt LMs for ASR, which means for each sentence in the testing data, we need to find a corresponding document for HDP, based on which topics are extracted and then LMs are dynamically adapted according to the topic information. Documents might have also been attached with corresponding roles. In the AMI Meeting Corpus, meetings are manually annotated with word transcription (in `*.words.xml`), whose time information were further obtained via forced alignment. Also available in the corpus are the segment annotations (in `*.segments.xml`). Role information for words can be easily determined from the annotations in the corpus. We used the procedure shown in Fig. 2 to obtain documents.

By collecting all documents for meetings belonging to the training and testing data respectively, we can obtain the training data for HDP model and the testing data for perplexity evaluation. The similar idea applies for dynamically finding documents for ASR experiments. The difference is that we do not have the segment annotations in this case. Instead speech segments, obtained by either automatical or manual approaches, are used as units for finding documents as well as for ASR. Notice in the ASR case we use an online unsupervised method,

```
foreach meeting m in the corpus
    retrieve words with time and role info for m;
    align all words in m to a common timeline;
    foreach segment s in m
      st = starttime(s); et = endtime(s)
      if et-st < winlen L: st = et-L;
      foreach w in words[st:et]
        if not stopword(w): doc(s) += w;
      end
      role(s) = role assigned to most words;
    end
  end
```

Fig. 2. The procedure used to define documents for the HDP/rHDP.

i.e., ASR hypotheses (with errors and time information) from previous segments are used to define documents for HDP inference for current segment. In both cases above, we simply ignore those segments without documents corresponding to them.

Incorporate Role Information. As a preliminary attempt, we consider the problem of introducing role into the HDP hierarchy for better topic modeling. In the scenario meetings of the AMI Meeting Corpus, each of the four participants in one meeting was assigned a different role (PM, ME, UI, or ID). Our intuitive idea for this is that, since different participants have different roles to play, there must be a different topic distribution, and in turn different dominant words, specific to each role. However, we still expect topic models to work as a whole on the corpus rather than four separate topic models. HDP is then the right choice, because it has a flexible framework to express DP dependencies using a tree structure.

To do this, documents were defined as described above for those scenario meetings with role information, a one-to-one mapping. We grouped the documents for each of the four roles, and assigned a DP G_{role} for each role, which then served as the parent DP in the HDP hierarchy (the base probability measure) for all DPs corresponding to documents belonging to that role. To share the topics among four roles, a global G_0 was used as the common base probability measure for the four role DPs G_{role}. See the graphical model shown in Fig. 1.(D) for detailed HDP hierarchy. Formally speaking, we used the following 3-level HDP, rHDP, to model topic and role information in the AMI Meeting Corpus:

$$G_0|\gamma, H \sim \text{DP}(\gamma, H), G_{\text{role}}|\alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0), G_j|\alpha_1, G_{\text{role}} \sim \text{DP}(\alpha_1, G_{\text{role}}) \quad (7)$$

Combine with n-gram. A topic in a HDP is a multinomial distribution over words in the vocabulary (denoted as $\boldsymbol{\phi}_k$), which, in this sense, can be considered as a unigram model. To be precise, we use $P_{\text{hdp}}(w|d)$ to denote the unigram

probabilities obtained by the HDP based on the jth document d. The HDP probability $P_{\text{hdp}}(w|d)$ is approximated as a sum over all the latent topics $\boldsymbol{\phi}_k$ for that document, supposing there are totally K topics in the HDP at the current time:

$$P_{\text{hdp}}(w|d) \approx \sum_{k=1}^{K} \phi_{kw} \cdot \theta_{dk} \tag{8}$$

where probability vector $\boldsymbol{\phi}_k$ is estimated during training and remains fixed in testing, while topic weights $\boldsymbol{\theta}_d|G_j \sim \text{DP}(\alpha_0, G_0)$ are document-dependent and thus are calculated dynamically for each document. For rHDP, the different is that the topic weights are derived from role DPs, i.e., $\boldsymbol{\theta}_d|G_j \sim \text{DP}(\alpha_1, G_{\text{role}})$.

As in [18], we treat $P_{\text{hdp}}(w|d)$ as a dynamic marginal and use the following equation to adapt the baseline n-gram model $P_{\text{back}}(w|h)$ to get an adapted n-gram $P_{\text{adapt}}(w|h)$, where $z(h)$ is a normalisation factor:

$$P_{\text{adapt}}(w|h) = \frac{\alpha(w)}{z(h)} \cdot P_{\text{back}}(w|h) \qquad \alpha(w) \approx \left(\frac{P_{\text{hdp}}(w|d)}{P_{\text{back}}(w)} \right)^{\mu} \tag{9}$$

4 Experiment and Result

We report some experimental results in this section. The HDP was implemented as an extension to the SRILM toolkit[3]. All baseline LMs used here were trained using SRILM, and the Nbest generation and rescoring were based on a modified tool from SRILM.

Since we considered the role information, which is only available in scenario AMI meetings, we used part of the AMI Meeting Corpus for our experiments. There are 138 scenario meetings in total, of which 118 were used for training and the other 20 for testing (about 11 hours). We used the algorithm introduced in Section 3 to extract the corresponding document for each of those sentences in both training and testing data. The average number of words in the resulting documents for window lengths of 10 and 20 seconds was 10 and 14 respectively. Data for n-gram LMs were obtained as usual for training and testing.

We initialized both HDP and rHDP models with 50 topics, and $\beta = 0.5$ for Equation 2. HDP/rHDP models were trained on documents of 10 seconds window length from the scenario AMI meetings with a fixed size vocabulary of 7,910 words, by the Markov Chain Monte Carlo (MCMC) sampling method. The concentration parameters were sampled using the auxiliary variable sample scheme in [9]. We used 3000 iterations to burn-in HDP/rHDP models.

4.1 Perplexity Experiment for LMs

In order to see the effect of an adapted LMs on perplexity, we trained three baseline LMs on three data respectively: the first one is AMI n-gram training data, the second is the Fisher data (fisher-03-p1+p2), the third is the Hub-4

[3] http://www.speech.sri.com/projects/srilm

Table 1. The perplexity results of HDP/rHDP-adapted LMs.

LMs	Baseline	HDP-adapted	rHDP-adapted
AMI	107.1	100.7	100.7
Fisher	228.3	176.5	176.4
Hub-4	316.4	248.9	248.8
AMI+Fisher+Hub-4	172.9	144.1	143.9

broadcast news data (hub4-lm96). A fourth LM was trained using all the above three datasets. All the four LMs were trained with standard parameters using SRILM: trigrams, cut-off value of 2 for trigram counts, modified Kneser-Ney smoothing, interpolated model. A common vocabulary with 56,168 words was used for the four LMs, which has 568 out-of-vocabulary (OOV) words for the AMI n-gram test data.

The trained HDP and rHDP models were used to adapt the above four baseline n-gram models respectively, using the formula in Equation 9 with $\mu = 0.5$. We note the different vocabularies used by HDP/rHDP models and n-gram models here. Only those words occurring in both the HDP/rHDP vocabulary and the n-gram vocabulary were scaled using Equation 9. Table 1 shows the perplexity results for adapted n-gram models. We can see both HDP- and rHDP-adapted LMs produced significant reduction in perplexity. However, the performance makes no difference regarding the dynamic marginals from whether HDP or rHDP.

4.2 ASR Experiment

Finally, we investigated the effectiveness of adapted LMs based on topic and role information from meetings on a practical large vocabulary ASR system. The AMIASR system [19] was used as the baseline system.

We began from the lattices for the whole AMI Meeting Corpus, generated by the AMIASR system using a trigram LM trained on a large set of data coming from Fisher, Hub4, Switchboard, webdata, and various meeting sources including AMI. We then generated 500-best lists from the lattices for each utterance. The reason why we used Nbest rescoring instead of lattice rescoring is because the baseline lattices were generated using a trigram LM.

We adapted two LMs (Fisher, and AMI+Fisher+Hub4) trained in Section 4.1 according to the topic information extracted by HDP/rHDP models based on the previous ASR outputs, using a moving document window with length of 10 seconds. The adapted LM was destroyed after it was used to rescore the current Nbest lists. Two adapted LMs together with the baseline LM were then used to rescore the Nbest lists with a common language model weight of 14 (the same as for lattice generation) and no word insertion penalty.

Table 2 shows the WER results. LMs adapted by HDP/rHDP both yield an absolute reduction of about 0.7% in WER. This reduction is significant using a

Table 2. The %WER results of HDP/rHDP-adapted LMs.

LMs	SUB	DEL	INS	WER
Fisher	22.7	11.4	5.8	39.9
AMI-1g-adapted	22.4	11.3	5.7	39.4
HDP-adapted	22.2	11.3	5.6	39.1
rHDP-adapted	22.3	11.3	5.6	39.2
AMI+Fisher+Hub4	21.6	11.1	5.4	38.2
AMI-1g-adapted	21.3	11.0	5.4	37.8
HDP-adapted	21.2	11.1	5.3	37.6
rHDP-adapted	21.2	11.1	5.3	37.5

matched-pair significant[4] test with $p < 10^{-15}$. However the HDP and the rHDP have no significant difference in the WER performance.

To further investigate the power of HDP/rHDP-adapted LMs, we trained a standard unigram, AMI-1g, on the AMI training data, which is the same data used for HDP/rHDP training. This unigram was trained using the same vocabulary of 7,910 words as that for HDP/rHDP training. We then used this unigram as dynamic marginal to adapt the baseline LMs, also using the formula in Equation 9. The "AMI-1g-adapted" lines in Table 2 shows the WER results. We see, although AMI-1g-adapted LMs have lower WERs than that of the baseline LMs, HDP/rHDP-adapted LMs still have better WER performances (with 0.2–0.3% absolute reduction) than AMI-1g-adapted. Significant testing indicates that both improvements for the HDP/rHDP are significant, with $p < 10^{-6}$.

5 Discussion and Future Work

In this paper, we successfully demonstrated the effectiveness of using the topic (and partly role) information to adapt LMs for ASR in meetings. The topics were automatically extracted by a nonparametric model called HDP, which is an efficient and flexible Bayesian framework for topic modeling. By defining the appropriate 'documents' for HDP models, we got significant reduction in both perplexity, and WER in the task of rescoring Nbest lists for about 11 hours of AMI meeting data.

To our understanding, the reasons for the significant improvements by adapted LMs based on the topic and role information via the HDP come from the following sources. First, the meeting corpus we worked on is a domain-specific corpus with limited vocabulary, especially for those scenario meetings, with some words quite dominant during the meeting. So if we could roughly estimate the 'topic', and scale those dominant words correctly, then it is promising to improve the performance for LMs. Second, HDP models can reasonably extract topics, particularly on this domain-specific AMI Meeting Corpus. One interesting result we

[4] http://www.icsi.berkeley.edu/speech/faq/signiftest.html

found is that different HDP/rHDP models, though trained using various different parameters, did not make significant difference in either perplexity or WER evaluation. By closely looking at the resulting topics, we found that some topics have very high probability for appearing in most of the HDP/rHDP models, regardless of the different training parameters. One characteristic of those topics is that the top words in them normally have very high frequency. Third, the sentence-by-sentence style LM adaption further contributes to the improvements, which has been demonstrated by the example of AMI-1g-adapted LMs in Table 2. Language models are dynamically adapted according to the changes of topics detected based on the previous recognized results. This can be intuitively understood as a situation where there are K unigram LMs, and we dynamically select one unigram to adapt the baseline LMs according to the context (topic). In this paper, however, both the number of unigram models K and the unigram selected for one certain time are automatically determined by HDP/rHDP. Although this is unsupervised adaptation, it is still better than LM adaptation using static LMs trained on reference data.

One the other hand, the rHDP did not prove to have better performances than HDP in either perplexity or WER. Our interpretation for this is that we did not explicitly use the role information for adapting LMs, instead, only use it as an additional DP level for sharing topics among different roles. But as mentioned above, based on the AMI Meeting Corpus, which has very limited domain and consequently limited vocabulary words, this will not causes much differences in the resulting topics, no matter whether HDP or rHDP is used for topic modeling. Despite this, including the role information in the hierarchical DP framework can give us some additional information, i.e., the topics proportions specified to each role . This implies some space for our further investigation into incorporating the role information into the hierarchical Bayesian framework for language modeling, i.e., sampling the role randomly for each document, empirically analysing the differences between HDP and rHDP, and explicitly using the role for language modeling. Another possibility for further investigation is about the prior parameter for Dirichlet distribution: does prior knowledge from language help to set this parameter? Finally, more ASR experiments to verify the consistence and significance of this framework on more meeting data, e.g., a 5-fold cross-validation on the AMI Meeting Corpus, would be informative.

Acknowledgement

We thank the AMI-ASR team for providing the baseline ASR system for experiments. This work is jointly supported by the Wolfson Microelectronics Scholarship and the European IST Programme Project FP6-033812 (AMIDA). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

1. Bilmes, J.A., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Proceedings of HLT/NACCL. (2003) 4–6
2. Xu, P., Emami, A., Jelinek, F.: Training connectionist models for the structured language model. In: Empirical Methods in Natural Language Processing, EMNLP'2003. (2003)
3. Wallach, H.M.: Topic modeling: Beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA (2006)
4. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research **3**(Feb) (2003) 1137–1155
5. Blitzer, J., Globerson, A., Pereira, F.: Distributed latent variable models of lexical co-occurrences. In: Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. (2005)
6. Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the Annual Meeting of the ACL. Volume 44. (2006)
7. Huang, S., Renals, S.: Hierarchical Pitman-Yor language models for ASR in meetings. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'07). (2007)
8. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. Language Resources and Evaluation Journal **41**(2) (2007) 181–190
9. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association **101**(476) (2006) 1566–1581
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3** (2003)
11. Mrva, D., Woodland, P.C.: Unsupervised language model adaptation for mandarin broadcast conversation transcription. In: Proceedings of Interspeech 2006, Pittsburgh, USA (2006)
12. Tam, Y.C., Schultz, T.: Unsupervised lm adaptation using latent semantic marginals. In: Proceedings of Interspeech 2006, Pittsburgh, USA (2006)
13. Hsu, B.J., Glass, J.: Style and topic language model adaptation using HMM-LDA. In: Proceedings of EMNLP 2006, Sydney, Australia (2006)
14. Teh, Y.W.: Dirichlet processes. Submitted to Encyclopedia of Machine Learning (2007)
15. Teh, Y.W., Kurihara, K., Welling, M.: Collapsed variational inference for HDP. In: Advances in Neural Information Processing Systems. Volume 20. (2008)
16. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Annals of Statistics **1**(2) (1973) 209–230
17. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica Sinica **4** (1994) 639–650
18. Kneser, R., Peters, J., Klakow, D.: Language model adaptation using dynamic marginals. In: Proceedings of Eurospeech, Rhodes (1997) 1971–1974
19. Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., Vepa, J., Wan, V.: The AMI system for the transcription of speech in meetings. In: Proceedings of ICASSP'07, Hawaii, USA (2007)