Unsupervised Language Model Adaptation Based on Topic and Role Information in Multiparty Meetings

Songfang Huang, Steve Renals

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

Abstract

We continue our previous work on the modeling of topic and role information from multiparty meetings using a hierarchical Dirichlet process (HDP), in the context of language model adaptation. In this paper we focus on three problems: 1) an empirical analysis of the HDP as a nonparametric topic model; 2) the mismatch problem of vocabularies of the baseline n-gram model and the HDP; and 3) an automatic speech recognition experiment to further verify the effectiveness of our adaptation framework. Experiments on a large meeting corpus of more than 70 hours speech data show consistent and significant improvements in terms of word error rate for language model adaptation based on the topic and role information.

Index Terms: language model, adaptation, topic model, hierarchical Dirichlet process, participant role

1. Introduction

In recent years there has been growing research interest in automatic speech recognition (ASR) for multiparty meetings, which is of essential importance for the subsequent meeting processing such as content analysis, summarisation, discourse analysis, and information retrieval. Meetings are spontaneous, conversational, and multimodal by nature. This makes the automatic transcription of speech in meetings a more challenging task than for read speech. State-of-the-art meeting ASR systems currently use the standard n-gram language model (LM), which approximates the history as the immediately proceeding $n-1$ words. The n-gram LMs for meetings are typically trained on a large amount of background (out-of-domain) data, together with a small amount of meeting (in-domain) data.

In meeting ASR systems, the in-domain data for the LM is relatively sparse with the comparison to the background data, and it is infeasible or time-consuming to collect sufficient in-domain data by transcribing the meeting archives. LM adaptation, which aims to alleviate the domain mismatch problem, therefore becomes increasingly important in ASR for meetings. In past years, various LM adaptation techniques have been proposed and studied. There are broadly two types of approaches: *supervised* or *unsupervised*. More recently, some work has been done in the area of adapting n-gram LMs based on topic knowledge for ASR on different domains, for example, broadcast news [1, 2], lecture recordings [3], and meetings [4]. All these work used probabilistic latent semantic analysis (pLSA) or latent Dirichlet allocation (LDA) for topic modeling, from which the unigram marginals were estimated and further used to scale the background LMs [5]. However, many conversational ASR systems currently still favour the standard LM adaptation approaches, such as model and count interpolation, [6, 7]. One reason for this is because in conversational meetings there is no obvious single linear stream of words, much less a well-defined document, for topic modeling.

We consider in this paper an unsupervised LM adaptation for ASR using a domain-specific meeting corpus — the AMI Meeting Corpus[1] [8] collected by the AMI project, which consists of 100 hours of multimodal meeting recordings with comprehensive annotations at a number of different levels. About 70% of the corpus was elicited using a design scenario, in which the participants play the roles of employees—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. The information we use for the LM adaptation comes from two multimodal cues in meetings: the *topic* and the participant *role*.

In our previous work [9], we have introduced the modeling framework for the topic and role information in meetings using a hierarchical Dirichlet process (HDP) [10], and demonstrated its effectiveness on a subset of the AMI Meeting Corpus in terms of perplexity and word error rate (WER). That work featured the use of the HDP for topic modeling in meetings, and the exploitation of a moving window over the word streams to dynamically extract topics from sequential meeting data using the HDP. This paper continues that work, by further addressing the following questions in the context of LM adaptation. First, an empirical analysis was carried out for the HDP — a nonparametric topic model — to see the modeling behaviors compared to LDA [11], another popular topic model often used for LM adaptation. Second, we investigated the vocabulary mismatch problem between the large-vocabulary ASR system and the topic model, due to the fact that normally only those content words are included for topic modeling. Third, we conducted a comprehensive 5-fold ASR experiment on the whole AMI scenario meeting corpus to further verify the consistence and scalability of the improvements.

2. Modeling Framework

2.1. Topic Modeling using HDP

In topic models, 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). LDA pioneered the use of Dirichlet distribution as the prior for topic distribution $\boldsymbol{\theta}_d$. Figure 1(A) depicts the graphical model for LDA. The generative process for words in each document is as

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

Accepted after peer review of full paper

September 22 – 26, Brisbane Australia

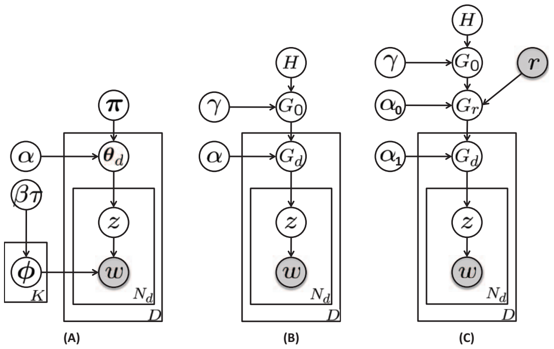


Figure 1: Graphical model depictions for (A) latent Dirichlet allocation (finite mixture model), (B) 2-level hierarchical Dirichlet process model, and (C) the role-HDP where G_r denotes the DP for one of the four roles (PM, ME, UI, and ID).

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, z_{id} the corresponding drawn topic, and Dirichlet priors are placed over the parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$, then

$$\begin{array}{ll} z_{id}|\boldsymbol{\theta}_d \sim \text{Mult}(\boldsymbol{\theta}_d) & w_{id}|z_{id}, \boldsymbol{\phi}_{z_{id}} \sim \text{Mult}(\boldsymbol{\phi}_{z_{id}}) \\ \boldsymbol{\theta}_d|\boldsymbol{\pi} \sim \text{Dir}(\alpha\boldsymbol{\pi}) & \boldsymbol{\phi}_k|\boldsymbol{\tau} \sim \text{Dir}(\beta\boldsymbol{\tau}) \end{array} \quad (1)$$

where $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$ are the corpus-wide distributions over topics and words respectively, and α and β are called the concentration parameters, controlling the amount of variability from $\boldsymbol{\theta}_d/\boldsymbol{\phi}_k$ to their prior means $\boldsymbol{\pi}/\boldsymbol{\tau}$.

In LDA, the number of topics K is determined in advance, i.e., $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_d$ are finite-dimensional vectors. The HDP, on the other hand, is a nonparametric extension to LDA, by using the stick-breaking construction [10] for $\boldsymbol{\pi}$ to accommodate a countably infinite number of topics, i.e., $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_d$ are now both infinite-dimensional vectors:

$$\pi'_k \sim \text{Beta}(1, \gamma) \qquad \pi_k = \pi'_k \prod_{l=1}^{k-1}(1 - \pi'_k) \quad (2)$$

A random measure defined as $G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\phi}_k}$ is then called a Dirichlet process (DP), with point masses located at $\boldsymbol{\phi}_k$. We write $G \sim \text{DP}(\gamma, H)$, with concentration parameter γ, base probability measure H, and $\boldsymbol{\phi}_k|H \sim \text{Dir}(\beta\boldsymbol{\tau})$. Reformulating topic modeling using the HDP according to [10], we have

$$G_0|\gamma, H \sim \text{DP}(\gamma, H) \quad G_d|\alpha, G_0 \sim \text{DP}(\alpha, G_0) \quad (3)$$

Figure 1(B) shows the corresponding 2-level HDP, which can be readily extended to as many levels as required.

The way we define a *document* in topic models is important, since it affects the scope of word co-occurrences to be considered for topic modeling. We used a moving window to define documents for the HDP: first align all words in a meeting along a common timeline; then for each sentence/segment, backtrace and collect those non-stop words belonging to a window of length L beginning from the end time of the sentence/segment.

2.2. Incorporate Role Information

We incorporate the participant role by extending the 2-level HDP in Figure 1(B) to a third level, as shown in Figure 1(C). An DP G_r is assigned for each of the four roles, which then served as the parent DP (the base probability measure) in the HDP hierarchy for all those DPs corresponding to documents belonging to that role. Formally speaking, we used the following 3-level HDP, rHDP, to model topic and role information:

$$G_0 \sim \text{DP}(\gamma, H), G_r \sim \text{DP}(\alpha_0, G_0), G_d \sim \text{DP}(\alpha_1, G_r) \quad (4)$$

2.3. Combine with n-gram LMs

As in [5], we use the dynamic unigram marginal from the HDP, $P_{\text{hdp}}(w|d)$, for LM adaptation:

$$P_{\text{adapt}}(w|h) = P_{\text{back}}(w|h) \cdot \left(\frac{P_{\text{hdp}}(w|d)}{P_{\text{back}}(w)}\right)^{\mu} / z(h) \quad (5)$$

where $P_{\text{back}}(w|h)$ is the baseline n-gram, $P_{\text{adapt}}(w|h)$ the adapted n-gram, $z(h)$ a normalisation factor, and $P_{\text{hdp}}(w|d) \approx \sum_{k=1}^{K} \phi_{kw} \cdot \theta_{dk}$ with $\boldsymbol{\phi}_k$ estimated during training and remaining fixed in testing, while $\boldsymbol{\theta}_d$ are document-dependent and thus are calculated dynamically for each test document.

2.4. Vocabulary Mismatch

There are 50k words in the vocabulary V_{asr} of our baseline LMs. After removing stop words in the AMI meeting corpus, we fix the size of vocabulary V_{hdp} as 7,910 words for our HDP/rHDP models. We get zero probabilities for $P_{\text{hdp}}(w|d)$ in (5) for those $w \notin V_{\text{asr}}$, which will be problematic for N-best rescoring. Therefore, we deal with this vocabulary mismatch problem in two ways: 1) *model interpolation*, in which for those $w \notin V_{\text{asr}}$ we directly assign the unigram probabilities from the background LMs to $P_{\text{hdp}}(w|d)$; and 2) *count interpolation*, in which $P_{\text{hdp}}(w|d) = \frac{C_{\text{back}}(w)+C_{\text{hdp}}(w)}{\sum_{w'}(C_{\text{back}}(w')+C_{\text{hdp}}(w'))}$ for each $w \in V_{\text{asr}}$. The second method corresponds to the MAP adaptation from the background unigram LMs for each topic by interpolating the count statistics from the background unigram $C_{\text{back}}(w)$ and the HDP $C_{\text{hdp}}(w)$ (normally boosted by some weights).

3. Experiment

3.1. Empirical Analysis

To empirically analyse the properties and behaviors of nonparametric models, we trained a set of HDP/rHDP models using various different parameters, for example, the initial number of topics ($k = 1, \ldots, 100$), the prior Dirichlet parameter for topics ($\beta = 0.5, 1.0$ in (1), and $\tau_w = 1/W$), and the length of document window ($L = 10, 20$ seconds). For LDA, the symmetric Dirichlet with parameters α_0/K was used for topic distribution $\boldsymbol{\theta}_d$. All models were trained using folds 2–4 of the AMI scenario meetings, with a fixed size vocabulary of 7,910 words, with the Markov Chain Monte Carlo (MCMC) sampling method. The concentration parameters were sampled using the auxiliary variable sample scheme in [10]. We ran 3,000 iterations to burn-in, then collected 10 samples from the posteriors to calculate the unigram perplexity on the fold-1 testing data, with a sample step of 5. Figure 2 shows the results, in which some random effects exist because they were based on only one run. We are interested in the following questions:

The comparison to LDA. The best number of topics K for LDA is around 10∼20. With appropriate values of k (i.e., $k = 5 - 50$), the HDP/rHDP can roughly converge to the best perplexity performance. However, for some extreme values of k (i.e., $k = 1, 100$), the HDP/rHDP failed to converge. This issue was caused by some local optima effects: from Figure 2 we can

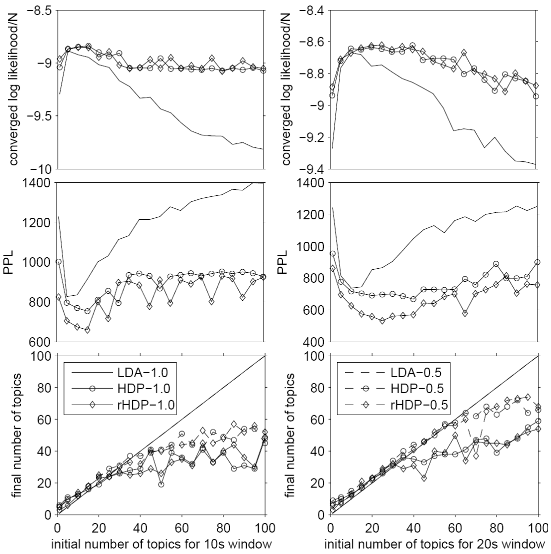
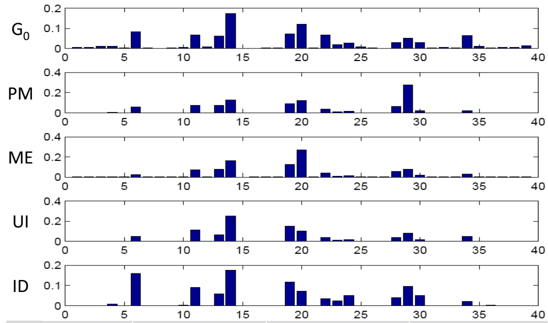


Figure 2: The empirical results of LDA and various HDP/rHDP models using different parameters, where the x-axis is the k for $L = 10$ and $L = 20$; the y-axis is: (top) the converged train log likelihood per word on folds 2–4, (middle) the perplexity on fold 1, and (bottom) K.



PM		ME		UI		ID	
(29) 0.28	(14) 0.13	(20) 0.27	(14) 0.16	(14) 0.25	(19) 0.15	(14) 0.17	(6) 0.16
DESIGN	BUTTON	REMOTE	BUTTON	BUTTON	RECOGNITION	BUTTON	CHIP
MEETING	BUTTONS	CONTROL	BUTTONS	BUTTONS	SPEECH	BUTTONS	SIMPLE
PROJECT	CHANNEL	LOOK	CHANNEL	CHANNEL	VOICE	CHANNEL	ADVANCED
MINUTES	SCREEN	MARKET	SCREEN	SCREEN	L_C_D_	SCREEN	REGULAR
USER	SCROLL	CONTROLS	SCROLL	SCROLL	SCREEN	SCROLL	REMOTE
INTERFACE	VOLUME	MEAN	VOLUME	VOLUME	MICROPHONE	VOLUME	L_C_D_
PRESENT	MENU	EASY	MENU	MENU	CONTROLLER	MENU	BATTERY
PRODUCT	L_C_D_	PRODUCT	L_C_D_	L_C_D_	REMOTE	L_C_D_	SPEAKER
DESIGNER	REMOTE	DESIGN	REMOTE	REMOTE	PROBLEM	REMOTE	EXPENSIVE
LOOK	WHEEL	BUTTONS	WHEEL	WHEEL	TECHNOLOGY	WHEEL	INFRARED
START	MEAN	FANCY	MEAN	MEAN	COFFEE	MEAN	STATION
SURE	CHANNELS	IMPORTANT	CHANNELS	CHANNELS	SPEAKER	CHANNELS	SAMPLE
BIT	CONTROL	PERCENT	CONTROL	CONTROL	COST	CONTROL	BATTERIES
GUESS	PRESS	FIND	PRESS	PRESS	CONTROL	PRESS	COST
THANK	KIND	LOT	KIND	KIND	SYSTEM	KIND	CONTROL

Figure 3: Examples of topic distributions for different roles, and top 2 topics (shown as top 15 words) for each role. This is based on the rHDP model with $k = 55, \beta = 0.5$, and $L = 10$.

see that the converged log training likelihood tends to a maximum when the perplexity is minimized. When we initialized k to extreme values, it got stuck at the local optima. Compared to the HDP/rHDP, however, this local optima effect is more severe for LDA. Therefore, the HDP/rHDP demonstrated its better modeling ability—seen in the perplexity results—and is more robust to local optima, by integrating over the topic cardinality.

The effect of role level. In terms of perplexity, we can see that the rHDP produced better results than the HDP. Moreover, the inclusion of role into the HDP provides some additional information. For example, we show in Figure 3 the four topic distributions specific to the four roles, and the top 3 example topics for each role from one rHDP model. We can see the rHDP reasonably captures the different topic distribution for each role. In this sense, the rHDP is a promising model for the inclusion of role into the HDP framework.

The initial number of topics k. Figure 2 shows that the HDP/rHDP added topics for $k = 1$, and pruned topics for $k = 100$. Both initializations can potentially converge to the best value of K . Due to the local optima effect, however, it is better for us to begin with a larger number of topics than to begin from smaller number of topics, i.e., $k = 100$ normally has lower perplexity comparing to $k = 1$ in our results.

The prior parameter β. The prior parameter β for the Dirichlet distribution plays an important role for the final value of K, with larger values of β leading to fewer final topics (see dash lines in Figure 2). Although for the HDP/rHDP we do not need to manually set the number of topics K as in LDA, it is necessary to take care when initializing the value for β.

The document window length L. The perplexity results for $L = 20$ are better, and more stable (with larger train likelihoods), than those for $L = 10$, with regard to different k. In addition, we found models with $L = 10$ suffered more severely from the local optima effect, for both LDA and HDP/rHDP. This suggests the local optima effect may be partly caused by the length of document window we used here. The word co-occurrence in these relatively short documents are sparse, which makes it hard for the models to escape from a local optima.

3.2. ASR Experiment

We used the AMI-ASR system [6] as the baseline. We began from the lattices for the whole AMI Meeting Corpus, generated by the AMI-ASR system using a trigram LM trained on a large set of data consisting of Fisher, Hub4, Switchboard, webdata, and various meeting sources including AMI, ICSI, NIST, and ISL. The baseline trigram LMs used for generating lattices in the AMI-ASR were adapted using model interpolation. We generated 500-best lists from the lattices for each utterance.

We selected parameters with $k = 25, \beta = 1.0$, and $L = 20$ to train rHDP models on each of the five folds of the AMI meeting data for the following ASR experiments. The topic information was extracted by the rHDP models based on the previous ASR outputs, using a moving document window with the length of 20 seconds. We used (5) to adapt the baseline LMs, with $\mu = 0.5$. Model interpolation (V1) and count interpolation (V2) were both used to deal with the vocabulary mismatch. The adapted LM was destroyed after it was used to rescore the current N-best lists. The rescoring used a common language model weight of 14 (the same as for lattice generation).

Table 1 shows the WER results. We found consistent WER reductions in all the 5-fold ASR experiments on the AMI Meeting Corpus, using LMs adapted by the rHDP. Although the absolute reductions are only about 0.2∼0.3% in WER, a significant testing using a matched-pair scheme[2] indicates that the reductions are all significant with $p < 0.01$. We also found that using count interpolation to deal with the vocabulary mismatch (V2) additionally provided a slightly better WER performance than the model interpolation version (V1).

ASR examples shown in Figure 4 illustrates the reasons for the improvements by adapting LMs based on the topic and

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

Table 1: The %WER results of rHDP-adapted LMs, where V1 and V2 denote the model and count interpolations respectively for dealing with the vocabulary mismatch.

FOLD	LM	SUB	DEL	INS	WER
1	baseline	20.7	11.1	5.2	37.0
	rHDP-V1-adapt	20.5	11.1	5.2	36.7
	rHDP-V2-adapt	20.2	11.8	4.6	36.6
2	baseline	19.6	11.0	4.9	35.5
	rHDP-V1-adapt	19.4	11.0	4.9	35.3
	rHDP-V2-adapt	19.1	11.7	4.4	35.2
3	baseline	20.7	11.1	4.8	36.6
	rHDP-V1-adapt	20.5	11.1	4.7	36.3
	rHDP-V2-adapt	20.2	11.8	4.2	36.3
4	baseline	19.3	10.9	5.3	35.5
	rHDP-V1-adapt	19.2	10.9	5.2	35.3
	rHDP-V2-adapt	18.9	11.6	4.7	35.2
5	baseline	23.1	12.4	6.1	41.6
	rHDP-V1-adapt	22.9	12.5	6.0	41.3
	rHDP-V2-adapt	22.5	13.1	5.3	41.0
all	baseline	20.6	11.3	5.2	37.1
	rHDP-V1-adapt	20.4	11.3	5.2	36.8
	rHDP-V2-adapt	20.1	12.0	4.6	36.7

role via the rHDP. 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/rHDP models can reasonably extract topics, particularly on this domain-specific AMI Meeting Corpus. Third, the sentence-by-sentence style LM adaption further contributes to the improvements. 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, based on which we dynamically estimate one interpolated unigram LM 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 the rHDP.

4. Conclusion

The conclusions we made in this paper are as follows: 1) from the empirical analysis, we believe the HDP overall is a powerful and flexible framework for topic modeling, attributed by its nonparametric property and hierarchical structure; 2) the HDP is sensitive to the initialization of k, because of the local optima effect. The local optima effect is partly affected by the way we define a document; 3) we are convinced that the unsupervised LM adaptation framework using the HDP for meeting ASR, as presented here, is effective, at least on the AMI Meeting Corpus; 4) for ASR, a HDP/rHDP model with lower empirical perplexity does not necessarily imply a lower WER. We observed WER results did not make much difference if we used a different HDP/rHDP model for LM adaptation; 5) it is important to define an appropriate document for the HDP in topic-based LM adaptation for meeting ASR; 6) a combination of LM adaptation approaches seems promising.

In future work, we will investigate the explicit use of participant role in meetings within the HDP for LM adaptation.

(A)	DOC	AGENDA TODAY DEFINE TALKED ENERGY KINETIC STUFF OPENING BATTERY COMPACT MINUTE PARTS MINUTES FUNCTIONAL DESIGN MEETING SIMPLE CHIP PRESENTATIONS DASH
	REF	***<s> OUR AGENDA WE'RE GOING TO DO AN OPENING I'M GOING TO REVIEW THE MINUTES OF </s>***
	BASE	<s> **RIGHT AND THEY WERE** GONNA DO THE OPENING A MINUTE OR IF YOU THE MINUTES OF </s>
	ADAPT	<s> **OUR AGENDA WE'RE** GONNA DO THE OPENING A MINUTE OR IF YOU THE MINUTES OF </s>
(B)	DOC	TEN TELETEXT BUTTONS NUMBERS BOSS AHEAD PAST PRETTY EASY AGES BUTTONS RECOGNITION FUNCTION REMOTE FINDING SCROLL CONTROL WHEEL REMOTE
	REF	***<s> YOU CAN INTRODUCE VOICE RECOGNITION BY UH FINDING BACK YOUR REMOTE </s>***
	BASE	<s> YOU CAN **AGES OF FOUR IS** RECOGNITION BY UH FINDING BACK YOUR REMOTE </s>
	ADAPT	<s> YOU CAN **AND USE THE VOICE** RECOGNITION BY UH FINDING BACK YOUR REMOTE </s>
(C)	DOC	ACTIVATE LIGHT LEADING CONSOLE STRONG BATTERY POWER KIND TECHNOLOGY EVALUATION FOCUSING CRITERIA L_ DUNNO C_D_ INSTRUCTIONS PANEL REQUIRES FEATURES HOUR L_C_D_ FORM LIGHT
	REF	***<s> WE COULD BECAUSE THE L. C. D. PANEL REQUIRES POWER AND THE L. C. D. IS A FORM OF A LIGHT </s>***
	BASE	<s> WE COULD BECAUSE THE L. C. D. PANEL REQUIRES **AN HOUR** AND THE L. C. D. IS A FORM OF LIGHT </s>
	ADAPT	<s> WE COULD BECAUSE THE L. C. D. PANEL REQUIRES **POWER** AND THE L. C. D. IS A FORM OF LIGHT </s>
(D)	DOC	NUMBERS CUSTOMISABLE CORNERS TABLE CHANNEL VOLUME FORTY VOLUME INTERESTED BRIGHTNESS CONTRAST READ SURE SHAPE REMOTE FRUIT BOWL IDEA STABLE CAT FEATURES
	REF	***<s> HAVE AN REMOTE IN THE SHAPE OF THE FRUIT OR A VEGETABLE OR WHAT EVER THEY LIKE </s>***
	BASE	<s> READ MORE IN THE SHAPE OF THE FRUIT BOWL OF **A STABLE** OR WHATEVER THE LIKE </s>
	ADAPT	<s> READ MORE IN THE SHAPE OF THE FRUIT BOWL OR **VEGETABLE** OR WHATEVER THEY LIKE </s>

(A)		(B)		(C)		(D)	
0.53	0.38	0.43	0.16	0.37	0.35	0.41	0.17
CHIP	MEETING	REMOTE	RECOGNITION	CHIP	REMOTE	BUTTON	FRUIT
BATTERY	DESIGN	CONTROL	SPEECH	BATTERY	SIGNAL	BUTTONS	BANANA
BATTERIES	MINUTES	LOOK	VOICE	BATTERIES	INFRARED	CHANNEL	SHAPE
ENERGY	PROJECT	FANCY	REMOTE	ENERGY	T_V_	SCREEN	COLOURS
SOLAR	USER	IMPORTANT	SAMPLE	SOLAR	BUTTON	SCROLL	SPONGY
KINETIC	INTERFACE	PERCENT	L_C_D_	KINETIC	CHIP	VOLUME	REMOTE
ADVANCED	PRESENTATION	USERS	SPEAKER	ADVANCED	BEEP	MENU	LOOK
SIMPLE	DESIGNER	CONTROLS	CONTROL	SIMPLE	LIGHT	L_C_D_	FEEL
STATION	THANK	EASY	SCREEN	STATION	CIRCUIT	WHEEL	VEGETABLES
REGULAR	START	TRUE	FIND	REGULAR	BOARD	PRESS	MEAN
L_C_D_	MARKETING	POINT	SENSOR	L_C_D_	INTERFACE	CHANNELS	VEGETABLE
POWER	INDUSTRIAL	FEEL	TECHNOLOGY	POWER	ACTUALLY	PUSH	FASHION
DOCKING	WORKING	BUTTONS	FEATURE	DOCKING	PRESS	FUNCTIONS	FRUITS
PRINT	THIRTY	INNOVATIVE	SIMPLE	PRINT	SEND	POWER	KIND
CELL	SURE	FIND	EASY	CELLS	USER	MEAN	CONTROL

Figure 4: Four ASR examples showing the rHDP-adapted LM works better than the baseline LM. DOC is the document formed from the previous ASR output and used to extract topics, with the top 2 showing at the bottom accordingly, REF is the reference, and BASE and ADAPT are the ASR hypotheses of the baseline LM and rHDP-adapted LM respectively.

5. Acknowledgements

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.

6. References

[1] D. Mrva and P. C. Woodland, "Unsupervised language model adaptation for mandarin broadcast conversation transcription," in *Proc. of Interspeech*, September 2006.

[2] Y.-C. Tam and T. Schultz, "Unsupervised lm adaptation using latent semantic marginals," in *Proc. of Interspeech*, Sep 2006.

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

[4] Y. Akita, Y. Nemoto, and T. Kawahara, "PLSA-based topic detection in meetings for adaptation of lexicon and language model," in *Proc. of Interspeech*, August 2007.

[5] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proc. of Eurospeech*, Rhodes, 1997.

[6] T. Hain and et al., "The AMI system for the transcription of speech in meetings," in *Proc. of ICASSP'07*, April 2007.

[7] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition," in *Proc. of ICASSP'07*, April 2007.

[8] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[9] S. Huang and S. Renals, "Modeling topic and role information in meetings using the hierarchical Dirichlet process," *Proc. of Machine Learning for Multimodal Interaction (MLMI'08)*, 2008.

[10] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.