Using Participant Role in Multiparty Meetings as Prior Knowledge for Nonparametric Topic Modeling

Songfang Huang S.F.HUANG@ED.AC.UK
Steve Renals S.RENALS@ED.AC.UK
The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom

Abstract

In this paper we introduce our attempts to incorporate the participant role information in multiparty meetings for document modeling using the hierarchical Dirichlet process. The perplexity and automatic speech recognition results demonstrate that the participant role information is a promising prior knowledge source to be combined with language models for automatic speech recognition and interaction modeling for multiparty meetings.

1. Introduction

In recent years there has been growing research interest in the 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. In this paper, we consider an improved language model (LM) in a state-of-the-art large vocabulary ASR system for meetings, based on the prior knowledge of participant roles. More specifically, we estimate the word distribution over the role of each participant, i.e., $P(w|r)$, and use this as unigram marginals to adapt a conventional n-gram LM.

The AMI and AMIDA (http://www.amiproject.org) projects are dedicated to the development of technologies to enhance the recognition and interpretation of interactions between people in multiparty meetings (Renals et al., 2007). The AMI Meeting Corpus collected by the AMI project 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, 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 intuition is that, since different participants play different roles, there may be a different word distribution, and in turn different dominant words, specific to each role. For example, we expect a project manager is more likely to speak words relating to the coordination of meetings, i.e., meeting, project, or present, while a user interface designer may favor words on interaction mediums like screen, voice, or speech.

Topic models have received much attention in the machine learning community, which follows the "bag-of-words" assumption, i.e., words in a document are exchangeable. In this paper we attempt to incorporate the participant role as prior knowledge in topic models, by assigning role information to exchangeable words in a document. This could be achieved within the flexible framework of topic models, by introducing an additional observed variable for the role into the graphical model. By assuming that each role has a mixture distribution over the latent topics, we could infer the topic distribution specific to each role. We could further estimate $P(w|r)$ for each role r by integrating out the latent topics. Moreover, incorporating the role in topic models enables not only the *document* modeling, but also the *interaction* modeling in meetings.

An alternative approach to modeling the relationship between the participant role information and lexical words is to directly estimate the conditional probability $P(w|r)$ based on the co-occurrence statistics of roles and words, using the maximum likelihood principle. As a comparison to the probabilistic topic models, we also introduce in this paper a deterministic approach to modeling roles and words, by regarding the role as an additional feature (factor) of lexical words in an MLE-based LM.

Appearing in *the Workshop on Prior Knowledge for Text and Language Processing* at ICML/UAI/COLT 2008, Helsinki, Finland, 2008.

2. Modeling Approaches

We consider two modeling approaches to the estimation of $P(w|r)$: one is a hierarchical Bayesian model using the hierarchical Dirichlet process (HDP) as the prior, and the other is a factored approach using the factored language model (FLM).

2.1. Hierarchical Bayesian Model

The hierarchical Dirichlet process (Teh et al., 2006) is a nonparametric generalization of latent Dirichlet allocation (LDA), which extends the standard LDA model to infinite and hierarchical topic modeling.

Conversational speech consists of sequences of utterances, which do not comprise well-defined documents. We used the following procedure to obtain documents: for each scenario meeting, first align all the words in it along a common timeline; then for each sentence/segment, collect those non-stop words belonging to a window of length L as the document, by backtracing from the end time of the sentence/segment. The role that has been assigned to the most of words in the window is selected as the role for that document. We use a moving window with $L = 20$ seconds over the sequences of words to obtain documents.

We incorporate the participant role by extending the 2-level HDP (Teh et al., 2006) in Figure 1(A) to a third level, as shown in Figure 1(B), role-HDP. An DP G_r is assigned for each of the four roles (PM,ME,UI,ID), 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.

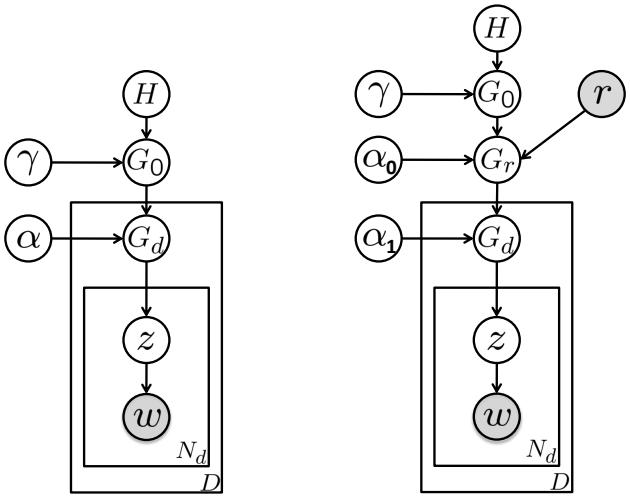

Figure 1. The graphical model depictions for (A) 2-level HDP, and (B) role-HDP.

2.2. Factored Language Model

One straightforward method for modeling words and roles is to use the maximum likelihood estimation based on the co-occurrences of words w and the role information r, i.e., training a bigram-like model $P(w|r) = \text{Count}(w, r)/\text{Count}(r)$. More generally, we can use a factored language model (Bilmes & Kirchhoff, 2003) to model words and role deterministically. The FLM, initially developed to address the language modeling problems faced by morphologically rich or inflected languages, is a generalization of standard n-gram language models, in which each word w_t is decomposed into a bundle of K word-related features (called *factors*), $w_t \equiv f_t^{1:K} = \{f_t^1, f_t^1, \ldots, f_t^K\}$. Factors may include the word itself. Each word in an FLM is dependent not only on a single stream of its preceding words, but also on additional parallel streams of factors. Combining with interpolation or generalized parallel backoff (GPB) (Bilmes & Kirchhoff, 2003) strategies, multiple backoff paths may be used simultaneously.

We exploit two factors for word w at time t: the word w_t itself and the corresponding role r_t, as shown in Figure 2. All the words in a sentence share a common role, i.e., $r_t = r_{t-1} = \ldots = r_1$ in Figure 2. We use a simple backoff strategy, for example, by moving from the model $P(w_t|r_t)$ directly down to the unigram model $P(w_t)$. We refer this model to the role-FLM.

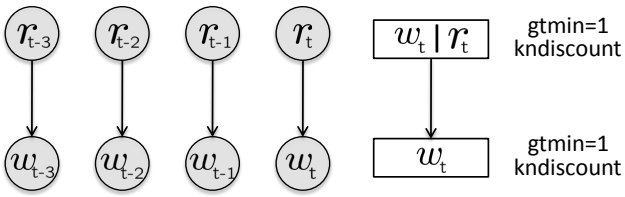

Figure 2. The graphical model representation and backoff path for role-FLM.

2.3. Combination with n-gram LMs

As in (Kneser et al., 1997), we use the dynamic unigram marginals $P(w|r)$, from either the role-HDP or the role-FLM, for LM adaptation:

$$P_{\text{adapt}}(w|h) = P_{\text{back}}(w|h) \cdot \left(\frac{P(w|r)}{P_{\text{back}}(w)} \right)^{\mu} / z(h) \quad (1)$$

where h is the history of w, $P_{\text{back}}(w|h)$ the baseline n-gram, $P_{\text{adapt}}(w|h)$ the adapted n-gram, and $z(h)$ a normalisation factor. For the role-HDP, $P(w|r) \approx \sum_{k=1}^{K} \phi_{kw} \cdot \theta_{dk}$ with $\boldsymbol{\phi}_k$ estimated during training and remaining fixed in testing, while $\boldsymbol{\theta}_d$ are document-dependent (and in turn are role-dependent because $\boldsymbol{\theta}_d$

are derived from G_r) and thus are calculated dynamically for each test document.

3. Experiments and Results[1]

3.1. Empirical Experiment

We first carried out some empirical analyses for the HDP and the role-HDP. The HDP was implemented as an extension to the SRILM toolkit[2]. We trained the HDP and the role-HDP models using different values ($k = 1, \ldots, 100$) for the initial number of topics. We used uniform distribution for H, i.e., $H_w = 1/W$. All models were trained using the fold-2−4 of the AMI scenario 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 (Teh et al., 2006). 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 the sample step of 5.

Figure 3 shows the perplexity results, from which we can see that the role-HDP produced better results than the HDP. Our understanding for the improvement is that by using the role prior knowledge, documents with the same role share the strengths in the HDP framework. In addition, we show in Figure 4 the top two topics for each role. It is interesting to find out that each role has some specific topics with high probabilities, while they also tend to interact with each other on some common topics, i.e., the button topic appears with high probability for all the four roles in Figure 4.

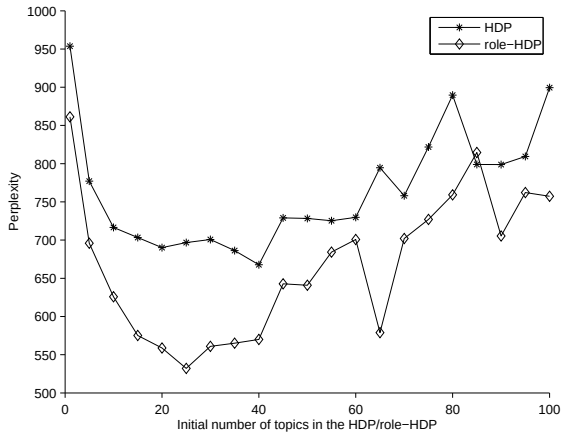

Figure 3. The perplexity results for the HDP/role-HDP.

[1]Some of the results here were appearing in another paper by the authors in (Huang & Renals, 2008).

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

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 4. The example topics for the four roles using the role-HDP.

3.2. ASR Experiment

To further investigate the effectiveness of employing the role as prior knowledge for topic modeling, we performed ASR experiments on multiparty 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). The procedure and parameters used to train the HDP/role-HDP were the same as those used in Section 3.1, except that we used a different split-up of the AMI scenario meetings for ASR.

We trained two baseline LMs: the first one used the Fisher conversational telephone speech data (fisher-03-p1+p2), and the second used three datasets from the AMI training data, the Fisher, and the Hub-4 broadcast news data (hub4-lm96). The two baseline 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 two LMs, which has 568 out-of-vocabulary (OOV) words for the AMI test data.

We investigated the effectiveness of the adapted LMs based on topic and role information from meetings on a practical large vocabulary ASR system. The AMI-ASR system (Hain & et al., 2007) 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. We adapted the two baseline LMs (Fisher and AMI+Fisher+Hub4) using Equation (1) according to the unigram marginals from the role-FLM, the HDP, and the role-HDP respectively. For the HDP, we used $P(w|d) \approx \sum_{k=1}^{K} \phi_{kw} \cdot \theta_{dk}$ as the unigram marginals, i.e., the difference from $P(w|r)$ by the role-HDP is that in the HDP $\boldsymbol{\theta}_d$ are only document-dependent

Table 1. The %WER results of ASR experiments using adapted LMs on the AMI scenario meetings.

LMs	SUB	DEL	INS	WER
Fisher	22.7	11.4	5.8	39.9
role-FLM-adapted	22.5	11.1	5.9	39.5
HDP-adapted	22.2	11.3	5.6	39.1
role-HDP-adapted	22.3	11.3	5.6	39.2
AMI+Fisher+Hub4	21.6	11.1	5.4	38.2
role-FLM-adapted	21.4	10.9	5.6	37.9
HDP-adapted	21.2	11.1	5.3	37.6
role-HDP-adapted	21.2	11.1	5.3	37.5

but not role-dependent. The topics were extracted by the HDP/role-HDP models based on the previous ASR outputs, using a moving document window with a length of 10 seconds. Three adapted LMs together with the baseline LM were then used to rescore the 500-best lists with a common language model weight of 14 (the same as for lattice generation) and no word insertion penalty. The adapted LM was destroyed after it was used to rescore the current N-best lists.

Table 1 shows the word error rate (WER) results. We can see that both the HDP and the role-HDP adapted LMs yield significant reductions ($p < 0.01$ according to a matched-pair significance test [3]) in WER, comparing to the baseline LMs. Although the role-FLM adapted LMs also reduce the WER, this deterministic approach is not as effective as the probabilistic topic modeling by introducing a latent variable – topic. However, there is no significant difference between the HDP and the role-HDP.

4. Discussion

Although the approach we used here to incorporate the role as prior knowledge for topic modeling is straightforward, the preliminary experiments demonstrated not only the better perplexity and WER results, but also the ability for modeling specific topic distributions for each role. This suggests some future work on the use of role information as prior knowledge is worth further investigation in the following aspects:

Probabilistic. The fact that we only assign one role for each document implies that we may lose some information, because there are potentially multiple roles for a document by using a moving window to obtain documents. Therefore, it is better to use a more probabilistic way, for example, each document is regarded as a multinomial distribution over roles, and each role a multinomial distribution over topics. Moreover, this helps to model the interactions between roles.

Observed vs. Latent. Depending on whether we treat the role variable as observed or latent, we can exploit the role as prior knowledge for topic modeling (as in this paper), or use other information to infer the role for each document. The latter is useful for modeling the human interactions in multiparty meetings.

Application. Even if we observed reductions in perplexity, it is not trivial to *transfer* the advantage of using the role prior knowledge for topic modeling to real applications such as on language modeling for automatic speech recognition in meetings. We observed no significant difference between the HDP and the role-HDP for ASR. We are interested in either a method of explicitly conditioning on the role for language modeling, or an approach to tightly combining topic models and n-gram models.

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

Acknowledgments

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

Bilmes, J. A., & Kirchhoff, K. (2003). Factored language models and generalized parallel backoff. *Proceedings of HLT/NACCL* (pp. 4–6).

Hain, T., & et al. (2007). The AMI system for the transcription of speech in meetings. *Proc. of ICASSP'07*.

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

Kneser, R., Peters, J., & Klakow, D. (1997). Language model adaptation using dynamic marginals. *Proc. of Eurospeech*. Rhodes.

Renals, S., Hain, T., & Bourlard, H. (2007). Recognition and interpretation of meetings: The AMI and AMIDA projects. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566–1581.