Unsupervised Latent Speaker Language Modeling

Yik-Cheung Tam and Paul Vozila

Nuance Communications Inc.
One Wayside Road, Burlington, MA 01803, USA
{wilson.tam,paul.vozila}@nuance.com

Abstract

In commercial speech applications, millions of speech utterances from the field are collected from millions of users, creating a challenge to best leverage the user data to enhance speech recognition performance. Motivated by an intuition that similar users may produce similar utterances, we propose a latent speaker model for unsupervised language modeling. Inspired by latent semantic analysis (LSA), an unsupervised method to extract latent topics from document corpora, we view the accumulated unsupervised text from a user as a document in the corpora. We employ latent Dirichlet-Tree allocation, a tree-based LSA, to organize the latent speakers in a tree hierarchy in an unsupervised fashion. During speaker adaptation, a new speaker model is adapted via a linear interpolation of the latent speaker models. On an in-house evaluation, the proposed method reduces the word error rates by 1.4% compared to a well-tuned baseline with speaker-independent and speaker-dependent adaptation. Compared to a competitive document clustering approach based on the exchange algorithm, our model yields slightly better recognition performance.

Index Terms: speaker topic modeling, language model adaptation

1. Introduction

Language model adaptation has been an active research area for automatic speech recognition. One popular approach is latent semantic analysis (LSA) which enables topical information of a context to be effectively incorporated into a background model to improve performance. LSA has been evolved from traditional singular value decomposition [1] to probabilistic approaches such as probabilistic latent semantic analysis [2, 3, 4], and latent Dirichlet allocation (LDA) [5, 6, 7]. These approaches usually train LSA models using supervised text (e.g. web articles). Moreover, the size of an article is usually sufficiently long so that topics in the articles are well captured. In some applications, however, obtaining supervised training text from a specific domain may be costly. Speech utterances can be short and independent, making topic adaptation difficult from the limited context.

In this paper, we explore speaker language modeling to address adaptation on short utterances. Unlike other efforts, by necessity, our approach relies exclusively on unsupervised text from a speech recognizer for speaker language modeling and adaptation. First, the accumulated text from a speaker is considered as a document, implicitly capturing the popular topics for that speaker. Intuitively, similar speakers produce similar utterances. We employ correlated N-gram LSA [8] to derive a set of correlated latent speakers in an unsupervised fashion. Informally, different topical words within a speaker document have mutual triggering effects via LSA. Knowing the current topics (e.g. finance) from preceding speaker data may help predict the most-likely future topics (e.g. technology) of the speaker. With the accumulated text of a speaker, we predict the language model interpolation weights per speaker. The speaker-specific interpolation weights are then used as part of the speaker adapted language model during subsequent speech recognition.

Works related to using speaker information include multi-speaker language modeling [9] which integrates the word usage of other speakers in a meeting for word prediction of a speaker via a word clustering approach. Probabilistic LSA [4] is employed to combine topic and speaker models for language model adaptation via unigram rescaling.

This paper is organized as follows: Section 2 gives a brief review of correlated N-gram LSA and speaker adaptation. Section 4 presents experimental results using our research system. Section 5 concludes our work.

2. Correlated N-gram LSA

Bigram LSA [10] attempts to relax the bag-of-word assumption in LSA that each word in a document is generated irrespective of its position in a document. Figure 1 shows the graphical representation of trigram LSA where the top node represents the prior distribution over the topic mixture weights and the middle layer represents the latent topic associated to each word at the bottom layer. The document generation procedure of N-gram LSA is similar to LDA except that the word history is considered to generate the next word:

1. Sample θ from a prior distribution $p(\theta)$

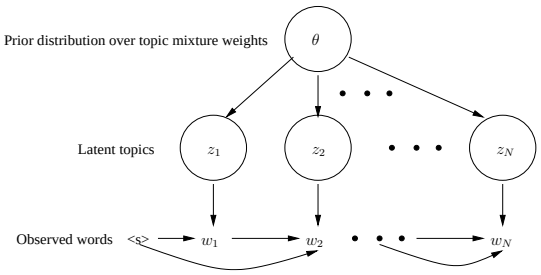

Figure 1: Graphical representation of trigram LSA.

2. For each word w_i at the i-th position of a document:

   (a) Sample topic label: $z_i \sim$ Multinomial(θ)

   (b) Sample w_i given the word history h_i and the topic label z_i: $w_i \sim p(\cdot|h_i, z_i)$

To model topic correlation, we use a Dirichlet tree as the prior distribution $p(\theta)$.

2.1. Model Training

We follow the same procedure in [8] for N-gram LSA training via variational Bayes inference. The joint likelihood of a document w_1^N, the latent topic sequence z_1^N and θ using N-gram LSA is written as follows:

$$p(w_1^N, z_1^N, \theta; \Lambda) \quad = \quad p(\theta) \cdot \prod_{i=1}^{N} p(z_i|\theta) \cdot p(w_i|h_i, z_i)$$

With a factorizable variational posterior distribution $q(z_1^N, \theta; \Gamma) = q(\theta) \cdot \prod_{i=1}^{N} q(z_i)$ over the latent variables, the lower bound of the marginalized document likelihood can be derived using the Jensen's inequality:

$$\begin{aligned}
& \log p(w_1^N; \Lambda) \\
= \quad & \log \int_\theta \sum_{z_1 \ldots z_N} q(z_1^N, \theta; \Gamma) \cdot \frac{p(w_1^N, z_1^N, \theta; \Lambda)}{q(z_1^N, \theta; \Gamma)} \\
\geq \quad & \int_\theta \sum_{z_1 \ldots z_N} q(z_1^N, \theta; \Gamma) \cdot \log \frac{p(w_1^N, z_1^N, \theta; \Lambda)}{q(z_1^N, \theta; \Gamma)} \\
= \quad & E_q[\log \frac{p(\theta)}{q(\theta)}] + \sum_{i=1}^{N} E_q[\log \frac{p(z_i|\theta)}{q(z_i)}] \\
& + \sum_{i=1}^{N} E_q[\log p(w_i|h_i, z_i)] = Q(w_1^N; \Lambda, \Gamma)
\end{aligned}$$

where the expectation is taken using $q(z_1^N, \theta; \Gamma)$. By partial differentiation on the auxiliary function $Q(\cdot)$ over the variational parameters Γ and setting the results to zero, we obtain the following E-step procedure (assuming LDA for simplicity):

E-steps:

$$\begin{aligned}
q(z_i = k) \quad & \propto \quad p(w_i|h_i, k) \cdot e^{E_q[\log \theta_k; \{\gamma_k\}]} \\
\gamma_k \quad & = \quad \alpha_k + \sum_{i=1}^{N} q(z_i = k) \\
\text{where } E_q[\log \theta_k] \quad & = \quad \sum_{k=1}^{K} \left(\Psi(\gamma_k) - \Psi(\sum_{k'} \gamma_{k'}) \right)
\end{aligned}$$

where $\{\alpha_k\}$ are the parameters of a Dirichlet prior. $\Psi(\cdot)$ denotes the derivative of the logarithm of the Gamma function.

For the M-step, we compute the partial derivative of the auxiliary function $Q(\cdot)$ over all training documents d with respect to the emission probability $p(v|h, k)$ and set the results to zero:

M-step: (unsmoothed)

$$\begin{aligned}
& p(v|h, k) \propto \sum_d \sum_{i=1}^{N_d} q(z_i = k|d) \cdot \delta(h_i, h)\delta(w_i, v) \\
= \quad & \frac{\sum_d C_d(h, v|k)}{\sum_d \sum_{v'=1}^{V} C_d(h, v'|k)} = \frac{C(h, v|k)}{\sum_{v'=1}^{V} C(h, v'|k)}
\end{aligned}$$

where N_d denote the number of words in document d and $\delta(x, y)$ is the Kronecker Delta function which sets to unity if x is equal to y. $C(h, v|k)$ denotes the fractional count of N-gram (h, v) belonging to topic k. Given the fractional N-gram counts, we could first rounding off the fractional counts to integers and apply any language model smoothing. Another approach is to apply fractional Kneser-Ney smoothing [8]. In this paper, we employ the former as approximation. To make the model training practical, we approximate $p(w_i|h_i, k)$ by $p(w_i|k)$ in the E-steps.

3. Unsupervised Speaker Adaptation

Using the accumulated unsupervised speaker text from a speech recognizer, we treat these as a speaker "document" and estimate the posterior of speaker s via the E-steps. The latent speaker model is obtained via linear interpolation:

$$\hat{\theta}_{sk} \quad = \quad \frac{\gamma_{sk}}{\sum_{k'} \gamma_{sk'}} \text{ for } k = 1 \ldots K \qquad (1)$$

$$p_{lsa}(v|h, s) \quad = \quad \sum_{k=1}^{K} \hat{\theta}_{sk} \cdot p(v|h, k) \qquad (2)$$

For new speakers without any accumulated text, we estimate the interpolation weights via simple averaging over the weights of training speakers:

$$\hat{\theta}_{*k} \quad = \quad \frac{1}{S} \sum_s \theta_{sk}$$

where S denotes the number of training speakers. As trivial baselines, we build speaker-independent (SI)

and speaker-dependent (SD) language models using the pooled data from all speakers, and the speaker-specific data respectively. Intuitively, the SI and SD models are the two extremes while the latent speaker model lies in-between them. The SI model is trained on more data but it may be too general. The SD model captures the speaker-specific behavior but it may lack sufficient training data. Finally, we interpolate the background language model with the SI, SD and the latent speaker models linearly:

$$\begin{aligned} p_a(v|h,s) &= \lambda_1 \cdot p_{bg}(v|h) + \lambda_2 \cdot p_{si}(v|h) \\ &\quad +\lambda_3 \cdot p_{sd}(v|h) + \lambda_4 \cdot p_{lsa}(v|h) \end{aligned}$$

where the interpolation weights are estimated using a combination of heuristics and grid search.

4. Experimental Setup

We evaluated the proposed speaker language modeling on an in-house speech recognition task using our research system. The speaker training corpus was a sample of 250K speakers totaling 8M words spanning a two-month period. This corpus consisted of unsupervised recognizer output. The test set was a sample of 1810 speakers totaling 57K words spanning the subsequent month. Manual transcriptions were used for accuracy assessment but not for adaptation. Most of the test utterances had less than ten words. The background model was an interpolated 4-gram LM built from diverse sources including large amounts of unsupervised in-domain data preceding the speaker training corpus. The background model was already optimized for the evaluation task. Language models were adapted for each test date using the speaker training corpus and all preceding test corpus from the recognizer output. Since both model training and adaptation used unsupervised text, the vocabulary size on all adaptation cases and the background model were equal and no out-of-vocabulary words were added. For speakers with non-empty accumulated data, we applied speaker-dependent adaptation. For latent speaker model, we used the training corpora to build an initial model. On each test date, we adapted the latent speaker models incrementally via folding in new preceding data followed by model update with ten EM iterations. Table 1 shows sample speaker-topics extracted from latent Dirichlet-Tree allocation. Finally, we interpolated the background model linearly with the adaptive components and tuned the interpolation weights accordingly. For comparison purpose, we employed the K-means style exchange algorithm [11] to perform "hard" clustering on speaker documents in contrast with "soft" clustering in latent Dirichlet-Tree allocation. In both cases, we used eight latent speakers for experiments with the same weight estimation procedure.

Top words of latent speakers
where need get want like buy
me for good taxi I'm looking
Oregon Portland Maine men Salem Washington
find California San_Francisco San_Jose
what where time weather playing closest

Table 1: Sample latent speakers from latent Dirichlet-Tree allocation.

LM Adaptation	Rel. PPLR	Rel. WERR
SI	31.4%	5.6%
SI+hard cluster	37.8	6.9
SI+LDA	38.9	8.0

Table 2: Oracle perplexity reduction (PPLR) and word error rate reduction (WERR) relative to the background model with various speaker adaptation approaches tuned on manual reference.

4.1. Oracle Results

As a sanity check, we evaluated the upper-bound performance via estimating the interpolation weights using the per-speaker manual reference on each test date and in all models including the background model. We applied unconstrained EM to estimate the weights until convergence. Table 2 shows the word perplexity and the word error rate compared to the background model. By simply pooling all the speaker data, speaker-independent adaptation yielded significant reduction in perplexity and word error rates by 31.4% and 5.6% respectively. With the utterance content changing over time, speaker-independent adaptation captures the changes effectively with preceding speaker data. Adding the latent speaker models either using "hard" clustering or LDA-based "soft" clustering further reduced the perplexity and word error rates. The LDA-based approach performed slightly better than the "hard" clustering approach on both perplexity and word error rate.

LM Adaptation	Rel. WERR (%)
SI	4.8%
SI+hard cluster	5.4
SI+LDA	5.7
SI+SD	5.9
SI+SD+hard cluster	6.7
SI+SD+LDA	7.2

Table 3: Word error rate reduction (WERR) relative to the background model with various speaker adaptation approaches tuned on preceding unsupervised text.

# utterance	SI+SD	SI+LDA	SI+SD+LDA
0	0.0%	0.46%	0.46%
≤ 1	-0.13	0.58	0.51
≤ 2	0.25	0.38	1.00
≤ 3	0.41	0.59	1.54
≤ 4	0.23	0.29	1.50
≤ 5	0.06	0.23	1.39

Table 4: Comparison between speaker-dependent and latent speaker adaptation on limited preceding adaptation utterances in word error rate reduction (WERR) relative to speaker-independent adaptation.

4.2. Recognition Results

To evaluate the adaptation methods in a more realistic setting, we then used the preceding accumulated data from the speech recognizer for weight estimation. Table 3 shows the word error rate on various adaptation scheme compared to the background model. We observed similar word error rate reduction using SI adaptation. The LDA-based latent speaker model yielded additional gain but the gain was significantly smaller than the gain observed in the oracle experiments. This shows that the quality of input adaptation data is crucial. Factors such as recognition errors, relevancy of the preceding data, and the amount of speaker-specific text may affect the accuracy of weight prediction on a specific test date in the latent speaker model.

Integrating SI and SD adaptation further brought down the word error rates by 5.9% relative to the background model. Combining all adaptation approaches yielded the best results with 7.2% relative reduction in word error rate. Results are statistically significant at 0.1% significance level with respect to applying both SI and SD adaptation. Similar to previous results, the LDA-based "soft" clustering produced slightly better results than the "hard" clustering using the exchange algorithm.

With insufficient adaptation data, speaker-dependent adaptation may not be effective. To verify this hypothesis, we compute the relative word error rate reduction with respect to speaker-independent adaptation using a subset of test speakers with limited number of preceding adaptation utterances. Table 4 shows the robustness of latent speaker adaptation with better recognition performance than speaker-dependent adaptation.

5. Conclusions

We have presented unsupervised latent speaker language modeling. Both model training and prediction use purely unsupervised text from a speech recognizer. With correlated N-gram LSA for latent speaker language modeling, we have shown significant improvement in recognition performance compared to a strong baseline using speaker-independent and speaker-dependent adaptation. Accurate estimation of interpolation weights have shown crucial but however governed by the intrinsic recognition errors and relevancy of the accumulated adaptation data of a speaker. The topics explored by a speaker may change over time during system usage. In the future, we will explore complementary input for better adaptation.

6. References

[1] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 88, no. 8, pp. 63–75, August 2000.

[2] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proceedings of Eurospeech*, Budapest, Hungary, September 1999, pp. 2167–2170.

[3] D. Mrva and P. C. Woodland, "A PLSA-based language model for conversational telephone speech," in *Proceedings of ICSLP*, Jeju Island, Korea, October 2004, pp. 2257–2260.

[4] Y. Akita and T. Kawahara, "Language model adaptation based on PLSA of topics and speakers," in *Proceedings of ICSLP*, Jeju Island, Korea, October 2004, pp. 1045–1048.

[5] Y. C. Tam and T. Schultz, "Language model adaptation using variational Bayes inference," in *Proceedings of Interspeech*, Lisbon, Portugal, September 2005, pp. 5–8.

[6] A. Heidel, H. Chang, and L. Lee, "Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007, pp. 2361–2364.

[7] J. T. Chien and C. H. Chueh, "Latent Dirichlet language model for speech recognition," in *IEEE Workshop on Spoken Language Technology*, Goa, India, December 2008, pp. 201–204.

[8] Y. C. Tam and T. Schultz, "Correlated bigram LSA for unsupervised LM adaptation," in *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2008.

[9] G. Ji and J. Bilmes, "Multi-speaker language modeling," in *Proceedings of HLT-NAACL*, Boston, USA, May 2004.

[10] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proceedings of ICML*. New York, NY, USA: ACM Press, 2006, pp. 977–984.

[11] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modeling," in *Proceedings of Eurospeech*, Berlin, Germany, September 1993.