Language Model Adaptation Using Latent Dirichlet Allocation and an Efficient Topic Inference Algorithm

Aaron Heidel[1], Hung-an Chang[2], and Lin-shan Lee[1]

[1]Dept. of Computer Science & Information Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

[2]Spoken Language Systems Group, CSAIL, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

aaron@speech.ee.ntu.edu.tw, hung_an@csail.mit.edu, lslee@gate.sinica.edu.tw

Abstract

We present an effort to perform topic mixture-based language model adaptation using latent Dirichlet allocation (LDA). We use probabilistic latent semantic analysis (PLSA) to automatically cluster a heterogeneous training corpus, and train an LDA model using the resultant topic-document assignments. Using this LDA model, we then construct topic-specific corpora at the utterance level for interpolation with a background language model during language model adaptation. We also present a novel iterative algorithm for LDA topic inference. Very encouraging results were obtained in preliminary experiments with broadcast news in Mandarin Chinese.

Index Terms: language model, unsupervised adaptation, topic modeling, speech recognition

1. Introduction

Statistical n-gram-based language models have long been used to estimate probabilities for the next word given the preceding word history. Although they have proven to work extremely well, researchers quickly noticed signs that their performance improvements given increasingly large corpora were beginning to asymptote [1]. Many attempts, therefore, have been made to compensate for their most notable weakness: no consideration of long distance dependencies, or no understanding of semantics.

One of the earliest such attempts introduced was the cache-based technique, which took advantage of the generally observed "burstiness" of words by increasing the probability of words in the history when predicting the next word [2]. This technique was then generalized using trigger pairs, in which the observation of certain "trigger" words increases the probability of seeing correlated words; furthermore, a maximum entropy approach was used to combine the probability estimates of multiple triggers into a reasonable joint estimate [3].

Another well-known approach is the sentence-level mixture model, which used topics identified from a heterogenous training corpus by automatic clustering [4]. Improvements were demonstrated in both perplexity and recognition accuracy over an unadapted trigram language model.

In this paper, we propose an improved language model adaptation scheme that utilizes topic information obtained from first-pass recognition results using a novel topic inference algorithm for latent Dirichlet allocation (LDA) [5]. A mix of language models is constructed that best models this topic information; that is, a set of significant topic-specific language models (topic LMs) is selected, and their corresponding weights when interpolated with a background LM are set. The resulting mix is then used to find an adapted hypothesis. The principle difference between this and a recently proposed approach [6] is that where they adapt the background LM according to the LDA-inferred unigram distribution, the proposed method decomposes the background LM into topic LMs using utterance-level n-gram counts. In fact, the proposed method is different from all such approaches that directly manipulate the background LM according to some unigram distribution based on the adaptation text. This approach is also conceptually simpler than a recent work using HMM-LDA [7], for example, in that no distinction is made between syntactic and semantic states.

In the next section, we describe the LDA framework for topic modeling and present the proposed inference algorithm, and in section 3 we outline the proposed language modeling scheme. Experimental results are presented in section 4, and our concluding remarks follow in section 5.

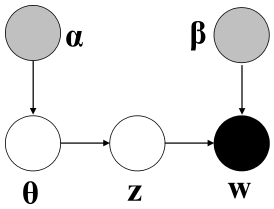

Figure 1: LDA topic sequence generation. The gray circles represent model parameters, while the white circles represent latent parameters.

2. Latent Dirichlet Allocation

2.1. LDA Model Representation

LDA [5] is a generative, probabilistic model characterized by the two sets of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\alpha} = [\alpha_1 \alpha_2 \cdots \alpha_k]$ represents the Dirichlet parameters for the k latent topics of the model, and $\boldsymbol{\beta}$ is a $k \times V$ matrix where each entry β_{ij} represents the unigram probability of the jth word in the V-word vocabulary under the ith latent topic.

Under the LDA modeling framework, the generative process for a document of N words can be summarized as follows. A vector for the topic mixture $\boldsymbol{\theta} = [\theta_1 \theta_2 \cdots \theta_k]$ is first drawn from the Dirichlet distribution with probability

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}, \qquad (1)$$

where $\Gamma(x)$ denotes the gamma function. Then, a topic sequence $\boldsymbol{z} = [z_1 z_2 \cdots z_N]$ is generated multinomially by $\boldsymbol{\theta}$ with $p(z_n = i|\boldsymbol{\theta}) = \theta_i$. Finally, each word w_n is chosen according to the probability $p(w_n|z_n, \boldsymbol{\beta})$, which is equal to β_{ij}, where $z_n = i$ and w_n is the jth word in the vocabulary. Figure 1 shows the generative process.

The joint distribution of a topic mixture $\boldsymbol{\theta}$, a set of N topics $\boldsymbol{z}$, and a set of N words $\boldsymbol{w}$ can thus be calculated by

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}). \qquad (2)$$

Integrating over $\boldsymbol{\theta}$ and summing over $\boldsymbol{z}$, we obtain the marginal distribution of $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \left(\sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \right) \right) d\boldsymbol{\theta}. \qquad (3)$$

2.2. LDA Topic Inference

We now illustrate how to infer the topic mixture $\boldsymbol{\theta}$ given a set of words $\boldsymbol{w}$. The inference requires the calculation of the marginal probability $p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. Unfortunately, the integration in eq. 3 is intractable due to the coupling between $\boldsymbol{\theta}$ and $\boldsymbol{w}$. Although a variational inference algorithm has been proposed to solve this problem [5], this algorithm still involves computationally expensive calculations such as the digamma function. To make topic inference faster and more efficient, we here propose an iterative algorithm to find the most suitable topic mixture $\hat{\boldsymbol{\theta}}$ for the set of words $\boldsymbol{w}$ under the mean square error (MSE) criterion.

The probability $p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ in eq. 3 can be further interpreted as an expectation over $\boldsymbol{\theta}$:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = E_{\boldsymbol{\theta}} \left[\prod_{n=1}^{N} \left(\sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \right) \right]. \qquad (4)$$

Under the MSE criterion, the most suitable mixture $\hat{\boldsymbol{\theta}}$ is the one that makes the probability of $\boldsymbol{w}$ equal to the expectation in eq. 4. That is, $\hat{\boldsymbol{\theta}}$ is chosen such that

$$\prod_{n=1}^{N} \left(\sum_{z_n} p(z_n|\hat{\boldsymbol{\theta}}) p(w_n|z_n, \boldsymbol{\beta}) \right) = E_{\boldsymbol{\theta}} \left[\prod_{n=1}^{N} \left(\sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \right) \right]. \qquad (5)$$

Because the Dirichlet distribution is continuous, we are guaranteed that such a $\hat{\boldsymbol{\theta}}$ exists. Hence, we can calculate $\hat{\boldsymbol{\theta}}$ using the following iteration.

First, we set the initial vector according to the prior information such that

$$\boldsymbol{\theta}^{[0]} = \left[\frac{\alpha_1}{\alpha_{sum}} \frac{\alpha_2}{\alpha_{sum}} \cdots \frac{\alpha_k}{\alpha_{sum}} \right], \qquad (6)$$

where $\alpha_{sum} = \sum_{i=1}^{k} \alpha_i$ and $\alpha_i / \alpha_{sum} = E[\theta_i|\boldsymbol{\alpha}]$. Let $\boldsymbol{\theta}^{[t]}$ denote the mixture after t iterations. The posterior probability λ_{ni} that word w_n is drawn from topic i can be derived by

$$\lambda_{ni} = \frac{\theta_i^{[t]} \beta_{iw_n}}{\sum_{j=1}^{k} \theta_j^{[t]} \beta_{jw_n}}, \qquad (7)$$

where $\theta_i^{[t]} \beta_{iw_n} = p(z_n = i \mid \boldsymbol{\theta}^{[t]}) p(w_n|z_n, \boldsymbol{\beta})$. Because each word is an independent and equally reliable observation under the LDA model, the posterior probability of each word has equal weight in determining the topic mixture. Thus, the new vector $\boldsymbol{\theta}^{[t+1]}$ can be calculated by

$$\theta_i^{[t+1]} = \frac{1}{N} \sum_{n=1}^{N} \lambda_{ni}. \qquad (8)$$

The iteration is continued until the mixture vector converges, resulting in the most suitable topic mixture for the word sequence $\boldsymbol{w}$.

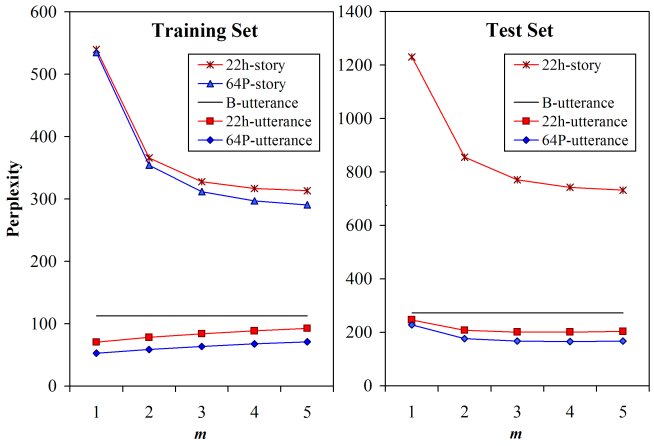

Figure 2: Perplexity for LDA-derived LM given topic mix size m. 'B' stands for the background trigram LM; 22 and 64 are topic counts; 'h' and 'P' stand for human- and PLSA-initiated; and 'utterance' and 'story' indicate per-utterance and per-story perplexities, respectively.

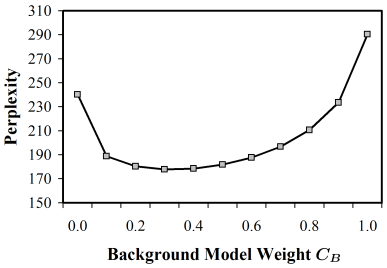

Figure 3: Perplexity given background model weight C_B when combined with top topic model.

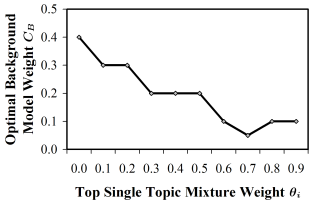

Figure 4: Optimal background model weight C_B given top topic mixture weight θ_i.

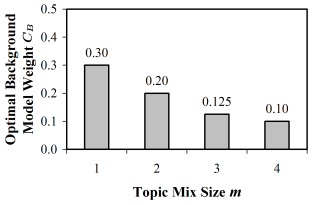

Figure 5: Optimal background model weight C_B given topic mix size m.

3. Language Modeling

3.1. Topic Corpus Construction and LM Training

Since LDA requires labelled training data, we first use PLSA to identify k latent topics in our training corpus. Given these topic labels, we train our LDA model, after which we proceed to assign each individual utterance in the training corpus to one of k topic corpora as follows: for each utterance, we infer the topic mixture $\boldsymbol{\theta}$ from which we choose the topic with the maximum weight, and append the utterance to this topic's corpus. We then use the resulting k topic-specific corpora to train each topic's trigram language model (the background trigram language model is trained on the entire training corpus). In our experiments, Good-Turing smoothing was used for all language models, and the SRI Language Modeling toolkit was used for all language model training and interpolation [8].

3.2. Language Model Adaptation

We perform utterance-level adaptation by inferring from each utterance u the LDA topic mixture $\boldsymbol{\theta}_u$ and interpolating the m topic LMs corresponding to the top m weights in $\boldsymbol{\theta}_u$ with the background LM; the background LM weight C_B is set to that which leads to the lowest overall perplexity on test data, and where θ_{u_i} is topic i's weight in $\boldsymbol{\theta}_u$, the interpolation weight C_{t_i} for topic i's LM is set to $C_{t_i} = (1 - C_B)\frac{\theta_{u_i}}{\sum_{j=1}^{m}\theta_{u_j}}$.

4. Experiments

For our training corpus, we used 20 months of text news supplied by Taiwan's Central News Agency (CNA), from January 2000 through August 2001. This corpus contains 245,417 stories, comprising 11,431,402 utterances (an utterance contains 5.8 words on average). For perplexity experiments, we used another 4 months of CNA text news from September through December 2001; this corpus contains 52,014 stories and 2,318,630 utterances. For recognition experiments, we used a random selection of 30 CNA-broadcast news stories from August and September 2002 comprising 261 utterances.

4.1. Topic LM Mixture Size Experiments

The results of our initial experiments are shown in Figure 2. Here a random set of 6000 and 5000 utterances was respectively selected from the training and testing corpora for perplexity evaluation. The experiment was conducted as follows: for each utterance u, the LDA topic mixture $\boldsymbol{\theta}_u$ was inferred and then per-utterance perplexity was calculated for interpolated mixtures of the the top m topics' language models, from $m = 1$ to $m = 5$. In these experiments, the background trigram LM was not included in the mixtures. The per-utterance perplexity of the entire set given the background trigram model was also calculated as 112.3 and 272.7 for training and testing, respectively. In addition, the experiment was conducted with two different sets of topics: 64 topics derived from those identified using PLSA, and 22 topics derived from those into which the Central News Agency categorized their news stories. Also, perplexities were calculated in a similar fashion but at the story level, to validate our utterance-based approach (the 64-topic PLSA-initiated per-story perplexity experiment was not run on the test set due to time constraints).

The results shown for the training set in Figure 2 seem to demonstrate the topical impurity of stories and the topical purity of utterances; that is, when topic mixtures are calculated by story, perplexities show a clear decreasing trend as the number of mixed-in language models (and thus the number of topics that are assumed to be represented by the story) increases. Conversely, when topic mixtures are calculated by utterance, perplexities show a clear rising trend trend as the number of mixed-in language models (and the number of topics that this utterance is assumed to represent) increases.

Results for the testing set, however, show not an increasing trend for per-utterance perplexities but a decreasing one: this could be evidence of overfitting. Still, though, there is a clear separation between per-story and per-utterance perplexities, as well as a large improvement for LDA-derived models with topic mixtures over the background model, and this improvement is

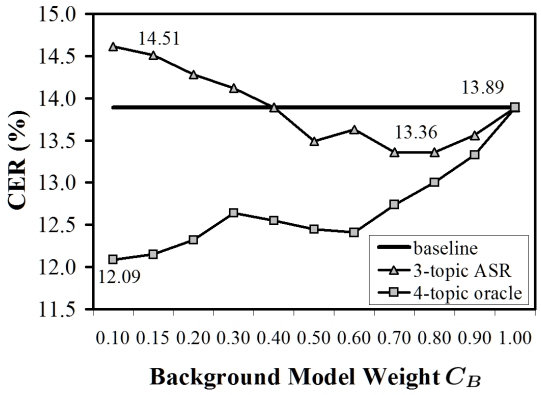

Figure 6: Recognition rates given background LM weight C_B.

consistent over both training and test sets.

In addition, the results shown for both sets seem to demonstrate the superiority of the 64 PLSA-initiated topics over the 22 human-initiated topics. The question here is whether this is an issue of PLSA-initiated versus human-initiated topics, or of topic granularity.

4.2. Interpolation Weight Experiments

Figures 3 through 5 show the results of experiments to find the optimal weights when interpolating the background model with the mix of topic models. Figure 3 shows detailed perplexity results when mixing the background model with the single topic with the highest weight, and indicates that a little background goes a long way; setting the background model weight C_B at 0.3 results in a 39% improvement in perplexity over just the background model alone ($C_B = 1.0$), and a 26% improvement over just the topic model ($C_B = 0$).

Figure 4 shows the variation of the optimal background model weight C_B with the top topic's mixture weight θ_i obtained from LDA as discussed in section 2. Here we can see that the top topic's mixture weight θ_i can be viewed as a measure of confidence, or alternately of utility, for this particular inference. When the topic's mixture weight θ_i is high, the background model is not needed as much as when the topic's mixture weight θ_i is low.

Finally, Figure 5 shows the optimal background model weight C_B for 1-, 2-, 3- and 4-topic mix sizes. The first bar for $m = 1$ is the result of $C_B = 0.3$ in Figure 3. The trend here is again that the more topics we add, the less we need the background model. Perplexity improvements over just using the background model in these cases ranged from 44% to 48%, although these results were obtained with a relatively small number of utterances and thus are less statistically trustworthy than the results for the single topic case.

4.3. Speech Recognition Experiments

Figure 6 shows the best results (measured in character error rate) of initial recognition tests in a two-pass framework that were performed on a development set of 261 utterances. First-pass results were used to infer topics, which were then used as the basis for language model adaptation in the second pass, as described as section 3.2. The 4-topic oracle experiment, in which reference transcripts were used for topic inference, yielded a 12.09% CER, a 13.0% relative improvement over the baseline (13.89% CER). However, for the 3-topic ASR experiment, where erroneous hypotheses were used for topic inference, we see an obvious mismatch between the optimal C_B as predicted by the perplexity experiments described in section 4.2 and that when performing actual recognition experiments. Here, setting C_B to the recommended 0.15 resulted in a 14.51% CER, a 4.5% relative degradation, whereas setting C_B to 0.7 or 0.8 resulted in a 13.36% CER, a 3.8% relative improvement.

We attribute these results to the effect of ASR errors distorting the results of topic inference. Intuitively, when topic inference are distorted, we should rely more on the background LM and less on the topic LMs dictated by the topic inference. Thus it makes perfect sense that while oracle experiments show the best results at more aggressive (lower) C_B values, real ASR experiments show better results when using more conservative (higher) C_B values. These experiments show great potential for improvements to ASR results if ways can be found to improve robustness when inferring topic mixtures.

5. Conclusions and Future Work

We have described a high-performance, unsupervised mechanism for topic mixture-based language model adaptation using LDA, and have proposed a novel decomposition of the training corpus for fine-grained topic LM estimation which results in improved perplexity and recognition accuracy. We have also presented a novel topic inference algorithm for LDA that is faster than the variational alternative.

In the future, our efforts will focus on improving robustness to ASR errors by performing topic inference on n-best lists instead of just 1-best hypotheses, by using adaptation segments larger than just one utterance, and also by finding run-time indicators for optimal C_B values instead of using a static C_B for all utterances.

6. References

[1] Ronald Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go From Here?," in *Proceedings of the IEEE*, 2000, pp. 1270–1278.

[2] R. Kuhn and R. De Mori, "A Cache-Based Natural Language Model from Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 570–583, 1990.

[3] Ronald Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer, Speech and Language*, pp. 187–228, 1996.

[4] R. Iyer and M. Ostendorf, "Modeling Long Distance Dependency in Language: Topic Mixtures vs. Dynamic Cache Models," in *Proceedings of ICSLP*, 1996, pp. 236–239.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, pp. 993–1022, 2003.

[6] Yik-Cheung Tam and Tanya Shultz, "Unsupervised Language Model Adaptation Using Latent Semantic Marginals," in *Proceedings of ICSLP*, 2006, pp. 2206–2209.

[7] B. J. Hsu and J. Glass, "Style & Topic Language Model Adaptation Using HMM-LDA," in *EMNLP*, 2006.

[8] Andreas Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of ICSLP*, 2002, pp. 901–904.