Integrating MAP, Marginals, and Unsupervised Language Model Adaptation

Wen Wang, Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA

{wwang,stolcke}@speech.sri.com

Abstract

We investigate the integration of various language model adaptation approaches for a cross-genre adaptation task to improve Mandarin ASR system performance on a recently introduced new genre, broadcast conversation (BC). Various language model adaptation strategies are investigated and their efficacies are evaluated based on ASR performance, including unsupervised language model adaptation from ASR transcripts and ways to integrate supervised Maximum A Posteriori (MAP) and marginal adaptation within the unsupervised adaptation framework. We found that by effectively combining these adaptation approaches, we can achieve as much as 1.3% absolute gain (6% relative) on the final recognition error rate in the BC genre.

Index Terms: automatic speech recognition, language model adaptation, supervised adaptation, unsupervised adaptation, broadcast conversation

1. Introduction

Building a large vocabulary automatic speech recognition (ASR) system for a new domain is a difficult task. One major reason is the fact that collection of the training data for acoustic and language model development is expensive and very labor-intensive. Acoustic model (AM) and language model (LM) adaptation attempt to adjust parameters of a well-trained existing model that is out-of-domain to the new domain, by exploring statistics obtained from a small in-domain seed corpus. There has been much effort recently in LM adaptation approaches for improving ASR performance; a wide range of LM adaptation approaches is reviewed in [1].

In this paper, we investigated LM adaptation approaches for a Mandarin large vocabulary ASR system originally developed to transcribe Mandarin broadcast news (BN). The new goal is to recognize Mandarin broadcast conversation (BC). These two genres are significantly mismatched. The BN genre follows its tradition and consists of "talking head" style broadcasts, i.e., generally one person reading a news script. The BC genre, by contrast, is more interactive and spontaneous, consisting of talk shows, interviews, call-in programs and roundtables. It is important to note that the classification of a program as BN or BC is a non-trivial task. For its data collection efforts, the Linguistic Data Consortium (LDC) defined the genre of a show to reflect the dominant genre of the program. Still, both genres could occur within a single program without indication of which portion is BN and which is BC. For BN we have abundant training data (1 billion words) to generate a well-trained LM. However, much less data (millions of words) is available to train a BC language model. Hence, this paper aims at exploring LM adaptation techniques to leverage the well-trained BN LM for the under-trained BC genre and improve ASR accuracy on BC data, within the setup of a single Mandarin ASR system.

Unsupervised LM adaptation has been extensively investigated in [2] by using ASR output transcripts. One of the observations is that count merge adaptation works better than model level interpolation adaptation, both producing gains on recognition error rate over the baseline using no adaptation. Work on supervised LM adaptation [3, 4] shows the effectiveness on ASR performance by using model level interpolation and counts merging. Closely related work on LM adaptation for BC transcription is reported in [5], where it is found that unsupervised adaptation of model level interpolation helps to improve BC ASR accuracy. Both N-gram adaptation and probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) models perform similarly and the combination of a dynamic N-gram with PLSA and LDA produced an additional gain. Here we study the efficacy of MAP variations and marginal LM adaptation approaches in supervised and unsupervised manner and explore effective ways to combine supervised and unsupervised approaches.

2. The SRI-UW Mandarin ASR System

2.1. Recognition System

In this work, various LM adaptation approaches were tested within the decoding infrastructure of the SRI-UW Mandarin ASR system [6] developed under the DARPA GALE program[1]. A diagram of the system used in the experiments is shown in Figure 1.
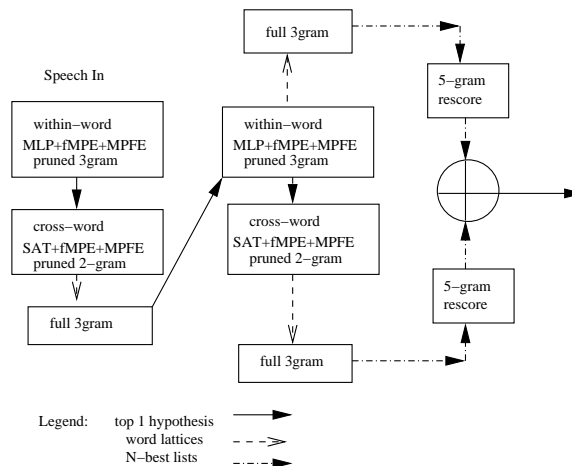

Figure 1: The SRI-UW Mandarin multi-pass ASR system.

---

[1]The goal of the GALE program is to develop computer software techniques to analyze, interpret, and distill information from speech and text in multiple languages.

Acoustic models were trained on 465 hours of audio data, including LDC Mandarin Hub4 training data, Mandarin TDT4 broadcast audio with closed captions filtered by flexible alignment [7], and data released in the first two quarters of the DARPA GALE program. Two sets of models were trained: one was a crossword triphone speaker-adaptive training (SAT) model based on MFCC+pitch front-end 42-dimensional features, trained using fMPE and MPE; the second model was similar except it used MFCC+pitch+MLP 74-dimensional features without SAT normalization, and within-word triphones. More details of the model training can be found in [6].

The LM training corpora comprised 1 billion words, including transcriptions of the acoustic training data, text corpora available under the GALE program, and 130M words of Web data. We used a unigram maximum likelihood based word segmentation algorithm [8] to segment the training text into multicharacter words. In the baseline system, two large 5-gram LMs were developed for N-best rescoring: a pruned, modified Kneser-Ney smoothed word 5-gram LM interpolated with two class-based 5-gram LMs, and an unpruned 5-gram LM smoothed by deleted interpolation.

As shown in Figure 1, the two acoustic models were each applied twice in four decoding passes in an interleaved way: after the first pass, the remaining three passes performed adapted decoding based on hypotheses generated from the previous pass. The last two passes generated N-best lists, which were rescored by the 5-gram LMs mentioned earlier. A character-level confusion network combination was then applied to the two sets of N-best lists to produce the final recognition results.

2.2. Language Model Training

Under the GALE program, there are two genres of broadcast audio, broadcast news (BN) and broadcast conversation (BC). Compared to the BN genre, BC is a recently introduced genre, with much less available training data. Hence, in this work, we focus on adapting the much better-trained BN-LM to the BC genre. Note that there may be overlap between BN and BC content in a particular audio file and LDC's classifications of a source program as BN or BC is meant to reflect the dominant genre. In this way, the transcripts released by LDC for the GALE program are labeled as belonging to the BN or BC genre. We used these labels to separate LM training data into two subsets, for training two genre-specific LMs, namely, BN-LM and BC-LM. In addition, we added the Mandarin conversational telephone speech (CTS) ASR system LM training data described in [9], including 1M word acoustic transcripts and 300M word Web downloaded conversational style and topic related data. The BN-LM training data sources hence include HUB4 1997 Mandarin BN acoustic transcripts, LDC Chinese TDT2, TDT3, TDT4 corpora, GALE Year 1 Quarter 1, 2, and Interim releases of BN audio transcripts, Multiple-Translation Chinese Corpus parts 1, 2, and 3, Chinese Gigaword corpus, Web downloaded BN transcripts and newswire text, and the CTS training data mentioned earlier. In total, BN-LM training data comprised 1,170M words. The BC-LM in-domain training data sources include GALE Year 1 Quarter 1, 2, and Interim releases of BC audio transcripts (3M words in total) and Web downloaded BC data (2M words in total), thus 5M words in total. Since the Web downloaded BC data is noisy, we used only the LDC released BC transcripts as the in-domain seed corpus for LM adaptation, denoted *BC-A*.

All LMs trained in this work used a 60K word vocabulary. The baseline BN-LM and BC-LM were trained using modified Kneser-Ney smoothing. The baseline LMs used in our ASR system were mixture models built by linearly interpolating the BN-LM and the BC-LM with weights optimized on a 100K-word heldout data set (denoted lm-tune) randomly sampled from the BN and BC training data, with 50K words each from BN and BC. Data sharing the same months with the dev06bn test set (a superset of the eval04 test set used in this work for testing BN decoding performance) and the same week with dev05bcm (the BC test set) were excluded. Interpolation weights are optimized to maximize the held-out data likelihood, using an expectation-maximization algorithm.

3. LM Adaptation Methods

We use $P_{\mathrm{BN}}(w|h)$ to denote the conditional probability of word w based on history h estimated for the BN-LM, $P_{\mathrm{BC}}(w|h)$ for the BC-LM, and $P_{\mathrm{BN-adapted}}(w|h)$ for the conditional probabilities according to the adapted BN-LM.

3.1. The Basic Unsupervised Language Model Adaptation

As described in Section 2.2, a baseline set of N-gram mixture LMs (for bigram, trigram, and 5-grams) was trained for the baseline ASR system. To perform the unsupervised language model adaptation, we first run the baseline ASR system to generate first-pass decoding hypotheses, and then optimize the linear interpolation weight λ between BN-LM and BC-LM based on the maximum likelihood criterion. The adapted mixture model probabilities are as follows:

$$P(w|h) = \lambda P_{\mathrm{BN}}(w|h) + (1 - \lambda)P_{\mathrm{BC}}(w|h) \qquad (1)$$

3.2. LM adaptation with unigram constraints

Kneser et al. [10] developed an approach to adapt the well-trained background LM (in our case, the BN-LM) as close as possible to the background estimates while constraining them to respect the unigram probabilities estimated from an in-domain seed corpus. That is, the unigram marginals are used to restrict the allowed adaptive N-grams:

$$\sum_h P_{\mathrm{BN-adapted}}(h) \cdot P_{\mathrm{BN-adapted}}(w|h) = \hat{P}_{BC-A}(w)$$

where $\hat{P}_{BC-A}(w)$ denotes the smoothed unigram probability estimated from the BC adaptation subset data BC-A. This adaptation method is an optimization procedure looking for

$$P_{\mathrm{BN-adapted}}(*|*) = \underset{P(*|*)}{\mathrm{argmin}} \sum_h P_{\mathrm{BN-adapted}}(h) \\ \cdot D(P(*|h)||P_{\mathrm{BN}}(*|h))$$

where D is relative entropy. After some approximations [10], the adapted LM is computed as

$$P_{\mathrm{BN-adapted}}(w|h) = \frac{\alpha(w)}{Z(h)} \cdot P_{\mathrm{BN}}(w|h)$$

where

$$\alpha(w) = \left(\frac{P_{BC-A}(w)}{P_{\mathrm{BN}}(w)} \right)^{\beta}$$

and $Z(h)$ is a normalization factor computed as

$$Z(h) = \sum_w \alpha(w) \cdot P_{\mathrm{BN}}(w|h)$$

The unigram marginals can be estimated from manual transcripts from in-domain data or from ASR recognition output. In this work, we investigated the effectiveness of both variants. We denote the former supervised unigram marginal adaptation and the latter unsupervised unigram marginal adaptation. In this work, β is chosen as 0.5.

3.3. MAP LM adaptation

3.3.1. MAP-1 adaptation

We investigated two variations on Maximum A Posteriori (MAP) adaptation in this work. In the first approach, denoted *MAP-1*, we explore the fact that the BN LM training data is in general heterogeneous and could be partitioned into different subsets based on their collection sources/domains. Each subset could make different contributions to the final adapted BN-LM based on their distributional similarity to the BC-A seed corpus. The procedure for MAP-1 adaptation can be represented as follows:

- Split the BN-LM training data into K subsets based on their collection sources (e.g., TDT, GALE-bn, newswire, etc.);
- Compute

$$C(w,h) = \sum_{i=1}^{K} \lambda_k(w,h) \cdot C_k(w,h)$$

where $C_k(w,h)$ is the counts of N-gram (w,h) appearing in the kth subset, and

$$\lambda_k(w,h) = \log\left(\frac{P_{\text{BC-A}}(w,h)}{P_k(w,h)}\right)^{\alpha}$$

- After counts are merged, the adapted BN-LM is estimated using a certain smoothing strategy.

Note that after this count merging method, the counts are not necessarily integers and hence any smoothing algorithm that relies on counts of counts (such as Kneser-Ney or Good-Turing) does not apply. Consequently, for MAP adaptation investigated in this paper, we apply the Witten-Bell smoothing in a backoff structure. In this approach, α is a smoothing constant that needs to be chosen empirically. We used a cross-validation based approach by first partitioning the BC-A data set into six equal-sized partitions and for each iteration, hold one partition as the test set and the rest as the adaptation data set (i.e., substituting BC-A in the above formula). We tested α values from 0 to 10 with step size 0.5. We combine counts and estimate LMs as illustrated above and test the perplexity of the adapted LM on the test set. We observed that with an increasing α, the perplexity on the adaptation data set always decreases, while the perplexity on the test set is a concave curve. The optimal α for each fold is hence selected as the value minimizing adapted LM perplexity on the test set and the final α for this task is the average of these fold-specific α values.

3.3.2. MAP-2 adaptation

We also investigated the more standard MAP adaptation approach, denoted *MAP-2*. Under this approach, N-gram counts from the BN training data and BC-A are combined and the adapted BN-LM is estimated as follows:

$$\hat{P}(w|h) = \frac{C_{\text{BN}}(h,w) + \tau C_{\text{BC-A}}(h,w)}{\sum_{w'} C_{\text{BN}}(h,w') + \tau \sum_{w'} C_{\text{BC-A}}(h,w')}$$

Then the final adapted BN-LM is computed using Witten-Bell smoothing on $\hat{P}(w|h)$ within a backoff structure. In this work, we chose a single τ value for all states h in the model, following [2].

4. Experimental Evaluation

The LM adaptation evaluations were conducted on a BC test set, dev05bcm (3 hours, 27K words, 5 shows) and a BN test set, eval04 (1.3 hours, 11K words, 3 shows). As shown in Figure 1, the first pass decoding used a within-word MFCC+MLP fMPE+MPE trained acoustic model and a pruned trigram LM, built as a mixture of BN-LM and BC-LM statically interpolated with weights optimized on lm-tune. In the unsupervised adaptation framework, the first pass decoding hypotheses were used to dynamically compute the interpolation weights between BN-LM and BC-LM and mixture models were rebuilt with the new weights (bigram and trigram LM for lattice generation and lattice expansion, and the 5-gram word LM used in N-best rescoring). Note that all 5-gram class-based LMs and the unpruned deleted-interpolation LM were not adapted in this work. We compared the effect of updating all mixture LMs and redecoding from scratch versus only using updated mixture 5-gram LMs during rescoring. The results are shown in Table 1.

As can be seen, unsupervised adaptation of BN-LM and BC-LM improved the final character error rate (CER) on bcdev05m by 0.7% absolute if all LMs were updated, while adapting only the rescoring 5-gram LM gave only 0.3% gain, suggesting that the earlier LM adaptation is performed, the more passes in the system can benefit from better hypotheses for acoustic model adaptation and better lattice generation, with consequently bigger impact on final system performance. Also, we observed that LM adaptation did not improve the CER on the BN test set, which is consistent with our hypothesis that the BN-LM is trained on a relatively sufficient amount of data and adaptation therefore does not improve it further. In the following experiments, we focus on examining the performance of LM adaptation techniques only on the BC test set. In Table 1, on the BC test set, we also observed that computing the dynamic interpolation weights at the show level instead of on the whole test set level produced a small gain, suggesting that even within one test set, shows vary as to their "BC-likeness".

Table 1: CERs for BC and BN test sets between adapting all LMs and redecoding from scratch versus adapting only the rescoring 5-gram LM. "Set" means that the dynamic interpolation weights were computed on the first pass decoding hypotheses of the whole test set, while "Show" means that the interpolation weights were computed on the first pass decoding hypotheses for each show in the test set.

Test Set	BN CER		BC CER	
	Set	Show	Set	Show
Baseline (static)	12.4		21.9	
Adapt 5-gram only	12.4	12.4	21.7	21.6
Adapt all LMs, redecode	12.5	12.4	21.5	21.2

Table 2 shows the results of applying unsupervised and supervised adaptation on the BN-LM using unigram marginals and then dynamically interpolating the adapted BN-LM and the BC-LM within the unsupervised adaptation framework. Note that all LMs were smoothed using the modified Kneser-Ney method. The "+" sign between two LMs in the first column of

the table indicates dynamic interpolation. By dynamically interpolating BC-LM and BN-LM and redecoding from scratch, the first pass CER is improved by 0.5% absolute and the final CER is improved by 0.7% absolute. Supervised adaptation of the BN-LM using unigram marginals uses the LDC-released BC transcripts as described in Section 2.2 (3M words in total) and the resulting adapted LM is denoted *BN-LM-UM-supervised*. Unsupervised adaptation of the BN-LM using unigram marginals uses the first pass decoding hypotheses and the resulting adapted LM is denoted *BN-LM-UM-unsupervised*. As can be seen, both supervised and unsupervised BN-LM adaptation using unigram marginals produced further performance improvements over the unsupervised adaptation of the unadapted BN-LM and BC-LM. Combination of first supervised unigram marginal adaptation and then dynamic mixing gave the largest CER improvement over the baseline (no adaptation): 0.9% absolute in the first-pass CER and 1.3% absolute in the final CER. The fact that supervised unigram marginal adaptation performs better than its unsupervised counterpart suggests that recognition errors can skew the unigram estimates and thus deteriorate the adapted LM.

Table 2: CERs for the BC test set using the unadapted BN-LM (denoted BN-LM) and unsupervised and supervised unigram marginal adapted BN-LM (denoted BN-LM-UM-unsupervised and BN-LM-UM-supervised), and the BC-LM within the unsupervised adaptation framework. The dynamic interpolation weights between adapted BN-LM and BC-LM were always computed at the show level. LMs were smoothed using modified Kneser-Ney.

Adaptation Setup	1st-pass	Final
Baseline (static)	24.9	21.9
BC-LM + BN-LM	24.4	21.2
BC-LM + BN-LM-UM-unsupervised	24.3	21.0
BC-LM + BN-LM-UM-supervised	**24.0**	**20.6**

In the third experiment, we investigated the efficacy of first applying supervised MAP adaptation of the BN-LM and then dynamically interpolating the adapted BN-LM and the BC-LM within the unsupervised adaptation framework. Note that as mentioned earlier, LMs here are all smoothed in a backoff structure using Witten-Bell discounting; hence, the baseline using no adaptation in Table 3 produced worse CER compared to the baseline using all modified Kneser-Ney smoothed LMs. The BN-LM was first MAP adapted on the LDC-released BC transcripts, resulting in BN-LM-MAP-1 and BN-LM-MAP-2 based on the two MAP variations described in Section 3, and then dynamically interpolated withe BC-LM based on the first pass decoding hypotheses from the baseline system. Overall the two MAP approaches produced comparable performance and a significant improvement in CER over the no-adaptation baseline: 1.4% absolute in the first-pass decoding CER and 1.6% absolute in the final CER. Since the modified Kneser-Ney smoothing algorithm could not be applied in this case, hence the final CERs with combined BC-LM and adapted BN-LM are not directly comparable to the best CERs in Table 2. We will explore ways to further aggregate the gains, for example, by conducting log-linear combination of all these LM scores.

In conclusion, we systematically investigate the effect of applying unsupervised language model adaptation and integrating MAP and marginal adaptation with the unsupervised adap-

Table 3: CERs for the BC test set using the unadapted BN-LM (denoted BN-LM) and supervised MAP adapted BN-LM, and BC-LM within the unsupervised adaptation framework. The dynamic interpolation weights between adapted BN-LM and BC-LM were always computed at the show level. LMs were smoothed using Witten-Bell discounting.

Adaptation Setup	1st-pass	Final
Baseline (static)	25.5	22.4
BC-LM + BN-LM	24.7	21.5
BC-LM + BN-LM-MAP-1	24.0	20.7
BC-LM + BN-LM-MAP-2	24.1	20.8

tation framework for the Mandarin BC recognition task. Unsupervised language model adaptation using BN-LM and BC-LM (trained on out-of-domain and in-domain data, respectively), achieves 0.7% absolute final CER reduction over the baseline CER 21.9% on the Mandarin BC test set. When first applying supervised adaptation using unigram marginals or MAP and then combining the adapted BN-LM and the BC-LM within the unsupervised adaptation framework, an additional improvement of 0.6%-0.8% absolute reduction in the final CER was achieved. We also observed that show-specific adaptation produced slightly better performance compared to adaptation on the whole test set. In future work, we will investigate the effect of the size of in-domain seed corpus on LM adaptation performance, as more and more BC data become available. Also, we will investigate the efficacy of other LM adaptation approaches and effective ways for their combinations.

5. Acknowledgments

6. References

[1] J. Bellegarda, "Statistical language model adaptation:review and perspectives," *Speech Communication*, vol. 42, no. 1, pp. 93–108, 2004.

[2] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proceedings of ICASSP*, 2003, pp. 224–227.

[3] P. Woodland, T. Hain, G. Moore, T. Niesler, D. Povey, A. Tuerk, and E. Whittaker, "The 1998 HTK broadcast news transcription system: Development and results," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[4] A. Ljolje, D. Hindle, M. Riley, and R. Sproat, "The AT&T LVCSR-2000 system," in *Proceedings of the NIST LVCSR Workshop*, 2000.

[5] D. Mrva and P. Woodland, "Unsupervised language model adaptation for Mandarin Broadcast Conversation transcription," in *Proceedings of International Conference on Spoken Language Processing*, 2006, pp. 1961–1964.

[6] J. Zheng, O. Cetin, M.-Y. Hwang, X. Lei, A. Stolcke, and N. Morgan, "Combining discriminative feature, transform, and model training for large vocabulary speech recognition," in *Proceedings of ICASSP*, Honolulu, Hawaii, 2007.

[7] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng, "An efficient repair procedure for quick transcriptions," in *Proceedings of International Conference on Spoken Language Processing*, Jeju, Korea, 2004, pp. 1961–1964.

[8] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin Broadcast News speech recognition," in *Proceedings of ICSLP*, Pittsburgh, 2006, pp. 1233–1236.

[9] T. Ng, M. Ostendorf, M. Hwang, M. Siu, I. Bulyko, and X. Lei, "Web-data augmented language models for Mandarin conversational speech recognition," in *Proceedings of ICASSP*, Philadelphia, 2005.

[10] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proceedings of Eurospeech*, Rhodes, 1997, pp. 1971–1974.