UNSUPERVISED LANGUAGE MODEL ADAPTATION

Michiel Bacchiani and Brian Roark

AT&T Labs-Research, 180 Park Ave., Florham Park, NJ 07932, USA
{michiel,roark}@research,.att.com

ABSTRACT

This paper investigates unsupervised language model adaptation, from ASR transcripts. N-gram counts from these transcripts can be used either to adapt an existing n-gram model or to build an n-gram model from scratch. Various experimental results are reported on a particular domain adaptation task, namely building a customer care application starting from a general voicemail transcription system. The experiments investigate the effectiveness of various adaptation strategies, including iterative adaptation and self-adaptation on the test data. They show an error rate reduction of 3.9% over the unadapted baseline performance, from 28% to 24.1%, using 17 hours of unsupervised adaptation material. This is 51% of the 7.7% adaptation gain obtained by supervised adaptation. Self-adaptation on the test data resulted in a 1.3% improvement over the baseline.

1. INTRODUCTION

Developing a speech recognition system for a new domain is costly, primarily due to the collection and preparation of the data required to train the system. Generally speaking, fairly large amounts of manually annotated data (tens of hours of data at a minimum for a large vocabulary system) are needed, which are very labor intensive to obtain.

Language model (LM) and acoustic model (AM) adaptation attempt to obtain models for a new domain with little training data, by leveraging existing ("out-of-domain") models. AM adaptation in particular has been studied extensively, both for application at test time [6, 3, 2], and for application on in-domain data other than the test set [5]. [5] showed that it is possible to obtain accurate acoustic models, using as little as 10 minutes of supervised training data to bootstrap a system, and then refining it, unsupervised, on an additional 135 hours.

In contrast to AM adaptation, LM adaptation has received much less attention. The most widespread approaches to supervised LM adaptation in a large vocabulary setting are model interpolation (e.g. [11]) and count mixing (e.g. [7]). Unsupervised LM adaptation has been investigated recently in [4, 9] by use of Automatic Speech Recognition (ASR) transcripts. [9] used the unweighted transcripts to build language models; [4] filtered or weighted based on confidence measures. The confidence annotation in that work was obtained from consensus hypothesis decoding [8]. In this paper, we study both the choice of the adaptation algorithm itself and the effectiveness of different LM adaptation strategies, i.e. varying the size and type of the adaptation sample and investigating the effect of iterative approaches.

Whereas in [5], a language model matching both the training and adaptation domains was available, and the focus was on AM adaptation, here we have an acoustic model which we believe to be well matched to both domains (since both are voicemail systems of the same type), and the focus is on LM adaptation. We present results for: LM adaptation on a sample from the domain of interest but different than the test sample; iterative LM adaptation; mixing supervised and unsupervised LM adaptation samples; and the use of unsupervised LM adaptation at test time (i.e. self-adaptation). All empirical results are presented with a multi-pass recognizer, using a variety of unsupervised speaker and channel normalization techniques, to evaluate if those adaptation gains are additive to the LM adaptation gains.

The LM adaptation algorithm is described in section 2. Experimental results, comparing the proposed algorithm to LM adaptation using supervised adaptation data and the effectiveness of the algorithm in a self-adaptation setting is described in section 3. Finally, section 4 discusses some conclusions that can be drawn from this work.

2. LANGUAGE MODEL ADAPTATION

Both count merging as well as model interpolation can both be viewed as a maximum *a posteriori* (MAP) adaptation [3] strategy with a different parameterization of the prior distribution. The model parameters θ are assumed to be a random vector in the space Θ. Given an observation sample $\mathbf{x}$, the MAP estimate is obtained as the mode of the posterior distribution of θ denoted as $g(. \mid \mathbf{x})$

$$\theta_{\text{MAP}} = \operatorname*{argmax}_{\theta} g(\theta \mid \mathbf{x}) = \operatorname*{argmax}_{\theta} f(\mathbf{x} \mid \theta) g(\theta). \tag{1}$$

The case of LM adaptation is very similar to MAP estimation of the mixture weights of a mixture distribution. In this case, the objective is to estimate probabilities for a discrete distribution across words, entirely analogous to the distribution across mixture components within a mixture density. Following the motivation and derivation in [3], a practical candidate for the prior distribution of the weights $\omega_1, \omega_2, \cdots, \omega_K$ is the Dirichlet density,

$$g(\omega_1, \omega_2, \cdots, \omega_K \mid \nu_1, \nu_2, \cdots, \nu_K) \propto \prod_{i=1}^{K} \omega_i^{\nu_i - 1} \tag{2}$$

where $\nu_i > 0$ are the parameters of the Dirichlet distribution. If the expected counts for the i-th component is denoted as c_i, the mode of the posterior distribution is obtained as

$$\hat{\omega}_i = \frac{(\nu_i - 1) + c_i}{\sum_{k=1}^{K} (\nu_k - 1) + \sum_{k=1}^{K} c_k} \quad 1 \leq i \leq K. \tag{3}$$

For a word w_i in n-gram history h, let the expected adaptation counts, in this application either from supervised transcripts or from ASR transcripts, be denoted as $\overline{c}(hw_i)$. Let the expected

0-7803-7663-3/03/$17.00 ©2003 IEEE

count for an n-gram history h be $\overline{c}(h) = \sum_i \overline{c}(hw_i)$. Let the corresponding expected counts from the out-of-domain sample be denoted as $\widetilde{c}(hw_i)$ and $\widetilde{c}(h)$. Let $\widetilde{c}_d(hw_i)$ and $\overline{c}_d(hw_i)$ denote the discounted counts for the out-of-domain and in-domain samples, respectively. Let $\widetilde{\mathrm{P}}(w_i \mid h)$ and $\overline{\mathrm{P}}(w_i \mid h)$ denote the probability of w_i in history h as estimated from the out-of-domain and in-domain samples, respectively. Then a count merging approach with mixing parameters α and β is obtained by choosing the parameters of the prior distribution for history h as $\nu_i = \widetilde{c}(h)\frac{\alpha}{\beta}\widetilde{\mathrm{P}}(w_i \mid h) + 1$ since in that case

$$\begin{aligned}\widehat{\mathrm{P}}(w_i \mid h) &= \frac{\widetilde{c}(h)\frac{\alpha}{\beta}\widetilde{\mathrm{P}}(w_i \mid h) + \overline{c}_d(hw_i)}{\sum_{k=1}^{K}\left[\widetilde{c}(h)\frac{\alpha}{\beta}\widetilde{\mathrm{P}}(w_k \mid h)\right] + \overline{c}(h)} \\ &= \frac{\alpha\widetilde{c}_d(hw_i) + \beta\overline{c}_d(hw_i)}{\alpha\widetilde{c}(h) + \beta\overline{c}(h)}. \qquad (4)\end{aligned}$$

On the other hand, if the parameters of the prior distribution for history h are chosen as $\overline{c}(h)\frac{\lambda}{1-\lambda}\widetilde{\mathrm{P}}(w_i \mid h) + 1$, the MAP estimate reduces to a model interpolation approach with parameter λ since in that case

$$\begin{aligned}\hat{\mathrm{P}}(w_i \mid h) &= \frac{\overline{c}(h)\frac{\lambda}{1-\lambda}\widetilde{\mathrm{P}}(w_i \mid h) + \overline{c}_d(hw_i)}{\sum_{k=1}^{K}\left[\overline{c}(h)\frac{\lambda}{1-\lambda}\widetilde{\mathrm{P}}(w_k \mid h)\right] + \overline{c}(h)} \\ &= \frac{\frac{\lambda}{1-\lambda}\widetilde{\mathrm{P}}(w_i \mid h) + \overline{\mathrm{P}}(w_i \mid h)}{\frac{\lambda}{1-\lambda} + 1} \\ &= \lambda\widetilde{\mathrm{P}}(w_i \mid h) + (1-\lambda)\overline{\mathrm{P}}(w_i \mid h). \qquad (5)\end{aligned}$$

3. EXPERIMENTAL RESULTS

We evaluated the unsupervised language model adaptation algorithm by measuring the transcription accuracy of an adapted voicemail transcription system on voicemail messages received at a customer care line of a telecommunications network center. The initial voicemail system, named Scanmail, was trained on general voicemail messages collected from the mailboxes of people at our research site in Florham Park, NJ. The target domain is also composed of voicemail messages, but for a mailbox that receives messages from customer care agents regarding network outages. In contrast to the general voicemail messages from the training corpus of the Scanmail system, the messages from the target domain, named SSNIFR, will be focused solely on network related problems. It contains frequent mention of various network related acronyms and trouble ticket numbers, rarely (if at all) found in the training corpus of the Scanmail system.

To evaluate the transcription accuracy, we used a multi-pass speech recognition system that employs various unsupervised speaker and channel normalization techniques. An initial search pass produces word-lattice output that is used as the grammar in subsequent search passes. The system is almost identical to the one described in detail in [1]. The main differences in terms of the acoustic model of the system are the use of linear discriminant analysis features; use of a 100 hour training set as opposed to a 60 hour training set; and the modeling of the speaker gender which in this system is identical to that described in [10]. Note that the acoustic model is appropriate for either domain as the messages are collected on a voicemail system of the same type. This parallels the experiments in [5], where the focus was on AM adaptation in the case where the LM was deemed appropriate for either domain.

System	FP	VTLN	CMA	MLLR
Baseline	32.7	30.0	28.3	28.0
In-domain	29.4	27.3	26.5	26.2
Count Merging	26.3	23.4	22.6	22.2
Interpolation	26.6	23.7	23.0	22.6

Table 1. Recognition performance using 3.7 hours of in-domain data for either training or adaptation using count merging or interpolation. The merging parameters were $\alpha = 1$ and $\beta = 5$, the interpolation parameter was $\lambda = 0.75$.

The language model of the Scanmail system is a Katz backoff trigram, trained on hand-transcribed messages of approximately 100 hours of voicemail (1 million words). The model contains 13460 unigram, 175777 bigram, and 495629 trigram probabilities. The lexicon of the Scanmail system contains 13460 words and was compiled from all the unique words found in the 100 hours of transcripts of the Scanmail training set.

For every experiment, we report the accuracy of the one-best transcripts obtained at 4 stages of the recognition process, after the first pass lattice construction (denoted as FP), after vocal tract length normalization and gender modeling (denoted as VTLN), after Constrained Model-space Adaptation (denoted as CMA) and after Maximum Likelihood Linear regression adaptation (denoted as MLLR).

For the SSNIFR domain we have available a 1 hour manually transcribed test set (10819 words) and approximately 17 hours of manually-transcribed adaptation data (163343 words). In all experiments, the vocabulary of the system is left unchanged. Generally, for a domain shift this can raise the error rate significantly due to an increase in the OOV rate. However, this increase in the experiments here is limited because the majority of the new domain-dependent vocabulary are acronyms which are covered by the Scanmail vocabulary through individual letters. The OOV rate of the SSNIFR test set, using the Scanmail vocabulary is 2%.

Table 1 lists the results obtained using 3.7 hours (38586 words) of manually transcribed SSNIFR domain data. The baseline result is the performance of the Scanmail system on the 1 hour SSNIFR test set without any adaptation. The in-domain result was obtained using a trigram language model trained on the 3.7 hours of in-domain data alone. The other lines give the performance of systems using the Scanmail language model, adapted with either count merging or interpolation. It can be seen that both adaptation approaches improve performance over the baseline (28.0%) and also improve over the in-domain trained model (26.2%). There is a larger improvement for the count merge adaptation than for the interpolation adaptation (5.8% vs. 5.4%). The count merging parameters ($\alpha = 1$ and $\beta = 5$) and interpolation parameter ($\lambda = 0.75$) were obtained empirically. Given these results, all subsequent experiments used a count merging approach with the same merging parameters.

Table 2 shows the results from supervised adaptation of the Scanmail language model using different sized subsets of the 17 hours of SSNIFR adaptation material. In these experiments, LM adaptation counts are obtained from the manual transcripts rather than from ASR transcripts. Table 3 repeats this experiment but in an unsupervised setting. Each subset of the adaptation data was first transcribed using an ASR system with the Scanmail language model. These transcripts were then used to obtain counts, and the

Fraction of the adaptation set (%)	FP	VTLN	CMA	MLLR
0	32.7	30.0	28.3	28.0
25	25.6	23.2	22.3	22.0
50	24.8	21.8	21.3	21.1
75	23.8	21.6	20.8	20.4
100	23.7	21.1	20.5	20.3

Table 2. Recognition on the 1 hour SSNIFR test set using systems obtained by supervised LM adaptation on various sized subsets of the 17 hour adaptation set.

Fraction of the adaptation set (%)	FP	VTLN	CMA	MLLR
0	32.7	30.0	28.3	28.0
25	28.9	27.0	25.8	25.5
50	28.4	26.0	25.2	24.8
75	28.1	25.6	24.9	24.7
100	28.2	25.6	24.9	24.6

Table 3. Recognition on the 1 hour SSNIFR test set using systems obtained by unsupervised LM adaptation on various sized subsets of the 17 hour adaptation set.

Scanmail language model was adapted using those counts. Although most of the improvement in accuracy comes from adapting on just 25% of the available 17 hours, improvements in both FP and MLLR accuracy were had by increasing the size of the adaptation sample.

Both supervised and unsupervised LM adaptation give performance improvements over the baseline using no adaptation. On a quarter of the 17 hour adaptation set, the unsupervised LM adaptation gives a 2.5% drop in the word error rate, compared to a 6.0% improvement using supervised LM adaptation. Increasing the amount of data used for LM adaptation to the full 17 hours gives an additional 1.7% and 0.9% improvement for the supervised and unsupervised cases respectively.

To investigate the effect of iterative LM adaptation, we used the system obtained by unsupervised LM adaptation on all of the 17 hour adaptation set to re-transcribe the entire adaptation set. We then used the counts from the MLLR-pass transcripts, together with the counts from the Scanmail language model, to obtain an adapted model. The results of adapted systems at multiple iterations are shown in table 4. A second iteration provided an additional 0.5% accuracy improvement. A third iteration gave no improvement in accuracy.

Iterations of adaptation	FP	VTLN	CMA	MLLR
0	32.7	30.0	28.3	28.0
1	28.2	25.6	24.9	24.6
2	27.9	25.1	24.4	24.1
3	28.0	25.3	24.7	24.3

Table 4. Recognition results of systems obtained by iterations of unsupervised LM adaptation using the entire 17 hour adaptation set. The adaptation counts were obtained from transcription with an adapted system.

Fraction of the adaptation set (%)	FP	VTLN	CMA	MLLR
50	25.4	22.2	21.7	21.5
100	25.0	22.1	21.5	21.3

Table 5. Recognition results of systems obtained by a second iteration of unsupervised LM adaptation using various sized subsets of the 17 hour adaptation set. The adaptation counts for 25% of the adaptation set are obtained from supervised transcripts, the rest from automatic transcription with an adapted system. That adapted system was obtained by supervised LM adaptation on 25% of the adaptation data. Hence the baseline is the 25% row of table 2.

To see to what extent the improvements of iterative LM adaptation are dependent on the starting point, we transcribed the adaptation set using the system obtained by supervised LM adaptation on 25% of the adaptation set. We then constructed adapted language models using the Scanmail model counts, the 25% supervised counts, and the counts obtained from the MLLR transcripts for the remaining subsets of the adaptation set. Both the supervised and unsupervised counts from the adaptation set were weighted with the same mixing parameter $\beta = 5$. The possibility of using different β parameters for the supervised and unsupervised counts was not investigated to allow a more direct comparison of the results of this mixed approach with that of the supervised-only results. However, given the difference in reliability of the supervised and unsupervised transcripts, it is possible that using multiple β parameters can result in improved accuracy. The results of the system adapted on the mixed supervised and unsupervised counts are shown in table 5. A comparison of these results with the performance of the system obtained just with supervised LM adaptation (table 2) demonstrates that using MLLR transcript-based counts in addition to to the supervised counts provides an additional accuracy improvement (21.3% vs. 22.0%) over using the supervised counts alone.

An alternative to an adaptation approach is to use the unsupervised counts obtained from ASR transcripts for model training directly. Table 6 shows the result using language models built from the MLLR transcripts of the adaptation set obtained by the baseline system. Using half of the adaptation set in this manner gave a 2% improvement in first-pass accuracy over the baseline; but this improvement is not additive, yielding just 0.4% improvement after all of the AM adaptation. The results do improve with more adaptation data: 2.9% FP accuracy improvement and 1.7% MLLR accuracy improvement.

Fraction of the adaptation set (%)	FP	VTLN	CMA	MLLR
50	30.7	28.4	27.7	27.6
100	29.8	27.0	26.4	26.3

Table 6. Recognition results of systems obtained by training language models solely from the transcripts produced by the baseline system on various subsets of the adaptation set.

A final LM adaptation scenario that was investigated is based on self-adaptation. In this scenario, the adaptation counts are obtained from the MLLR transcripts produced by the final search pass on the 1 hour test set. The test set is then re-transcribed using a

Initial model	FP	VTLN	CMA	MLLR
Scanmail	27.3	27.0	26.6	26.7
SSNIFR 17h unsup	25.5	24.5	24.1	24.0

Table 7. Recognition results of systems obtained by self-adaptation on the test set. Adaptation counts were obtained from the MLLR-pass test set transcripts produced by a system using the Scanmail or second iteration unsupervised adapted (see table 4) language models.

language model obtained by adaptation using the Scanmail counts and the adaptation counts from the test set. Table 7 shows the results from two such experiments. The experiments differed in the language model used for self-adaptation. In each experiment, the LM to be adapted was used to transcribe the test set. This LM was then adapted with the counts from the ASR transcript of the test set. One experiment used the baseline Scanmail language model; the other used the language model obtained by two iterations of unsupervised adaptation on the 17 hour adaptation set (see table 4). In both experiments, there is a large gain in the first-pass (FP) accuracy: 5.4% for the Scanmail trial (27.3% vs. 32.7%); 2.4% for the unsupervised adapted trial (25.5% vs. 27.9%). These gains, however, are not additive with the AM adaptation gains and reduce at the final search pass to 1.3% for the Scanmail trial (26.7% vs. 28.0%) and 0.1% for the unsupervised adapted trial (24.0% vs. 24.1%). This shows that the self-adaptation incorporates a part of the unsupervised AM adaptation gain. Self-adaptation does provide a gain in accuracy, but dependent on the starting point, since transcription accuracy improved for the baseline trial but not for the unsupervised adapted trial. The 1.3% improvement using self-adaptation alone on the baseline model is less than the 3.4% obtained by a single iteration of unsupervised adaptation on the 17 hour adaptation set.

4. CONCLUSIONS

This paper presents experimental results showing various approaches to unsupervised language model adaptation based on counts from ASR transcripts, each providing gains over the unadapted baseline system. Starting from a 28% word-error baseline, using 17 hours of in-domain adaptation data, unsupervised LM adaptation achieves 51% of the 7.7% adaptation gain obtained by supervised LM adaptation. A quarter of the 17 hour adaptation set, in a unsupervised setting, provides a 2.5% gain over the baseline: 64% of the gain obtained using the full 17 hours.

Iterative LM adaptation also improves accuracy, raising the accuracy gain from 3.4% to 3.9% with one additional iteration of the unsupervised adaptation approach. When starting with a model obtained by supervised adaptation on 25% of the adaptation set, iterative unsupervised adaptation still provides an additional improvement, raising the 6.0% gain from supervised adaptation by 0.7%.

Comparing the iterative unsupervised adaptation approach to a training approach, it shows that for a 17 hour adaptation sample, the gain from adaptation is 2.2% larger than that of training (3.9% vs. 1.7%).

Furthermore, self-adaptation on the 1 hour test set provides gains over the baseline system of 1.3%. This gain is, however, dependent on the starting point, since self-adaptation applied on top of an adapted model did not provide any additional gains. All self-adaptation experiments show a large improvement in the first-pass accuracy, however, this gain is not additive with the gains obtained from the AM normalization and self-adaptation algorithms. The LM adaptation process obtains part of the AM adaptation benefits by adapting on the transcripts that already have merited from AM adaptation.

One question not addressed in this paper is what the best adaptation strategy is when provided with a very large adaptation sample. Presumably, the advantage of LM adaptation over LM training will be reduced but it might also affect the effectiveness of iterative unsupervised approaches. Another question that is not addressed in this paper is the most beneficial approach to an approach that combines unsupervised acoustic and language model training. This raises the question of what level of system refinement can be obtained by combining unsupervised training approaches.

5. REFERENCES

[1] M. Bacchiani. Automatic transcription of voicemail at AT&T. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[2] M. J. F. Gales. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. *Computer Speech and Language*, pages 75–98, 1998.

[3] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[4] R. Gretter and G. Riccardi. On-line learning of language models with word error probability distributions. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 557–560, 2001.

[5] L. Lamel, J.-L. Gauvain, and G. Adda. Unsupervised acoustic model training. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 877–880, 2002.

[6] C. J. Legetter and P. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, pages 171–185, 1995.

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

[8] L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words: Lattice-based word error minimization. In *Proceedings of Eurospeech*, 1999.

[9] A. Stolcke. Error Modeling and Unsupervised Language Modeling. In *Proc. of the 2001 NIST Large Vocabulary Conversational Speech Recognition Workshop*, Linthicum, Maryland, May 2001.

[10] P. Woodland and T. Hain. The September 1998 HTK Hub 5E System. In *The Proceedings of the* 9^{th} *Hub-5 Conversational Speech Recognition Workshop*, 1998.

[11] 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 *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.