UNSUPERVISED LANGUAGE MODEL ADAPTATION FOR MEETING RECOGNITION

Gokhan Tur Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, 94025, USA
{gokhan,stolcke}@speech.sri.com

ABSTRACT

We present an application of unsupervised language model (LM) adaptation to meeting recognition, in a scenario where sequences of multiparty meetings on related topics are to be recognized, but no prior in-domain data for LM training is available. The recognizer LMs are adapted according to the recognition output on temporally preceding meetings, either in speaker-dependent or speaker-independent mode. Model adaptation is carried out by interpolating the n-gram probabilities of a large generic LM with those of a small LM estimated from the adaptation data, and minimizing perplexity on the automatic transcripts of a separate meeting set, also previously recognized. The adapted LMs yield about 5-9% relative reduction in word error compared to the baseline. This improvement is about half of what can be achieved with supervised adaptation, i.e., using human-generated speech transcripts.

Index Terms— speech processing, language modeling, meeting recognition, unsupervised adaptation

1. INTRODUCTION

One promising application of automatic speech recognition (ASR) is the automatic recognition (and eventually, understanding) of meetings. In many organizations, staff spend many hours each week in meetings; consequently, automatic means of transcribing, indexing, and summarizing meetings would greatly increase productivity of both meeting participants and non-participants. The annual NIST meeting recognition evaluations have become a driving force for research in meeting transcription technology, with substantial performance improvements in recent years [1]. In order to promote robustness and domain-independence, the NIST evaluations cover a range of meeting genres and topics, from largely open-ended, interactive chit-chat, to topic-focused project meetings, to technical seminars dominated by lecture-style presentations.

In real-life applications, meetings are often highly specialized in that they are concerned with technical topics or the specifics of the organization they occur in. Consequently, one cannot hope to have a recognition system that is well-matched to the topic or interaction style at hand. On the plus side, however, most meetings have a significant history of prior meetings on the same or related topics, and using similar settings and re-occurring speakers. This suggests that meeting recognition (in real-life settings) presents an ideal task for unsupervised adaptation of the recognition system. Adaptation is unsupervised because we assume that human annotation (e.g., transcription) of past meetings is not feasible or cost-effective, and therefore the recognition system has to rely only on its own output and available side information (such as speaker identities).

In this paper we focus on LM adaptation in a state-of-the-art meeting recognition system, using data from the CALO Meeting Assistant (CALO-MA) project. CALO-MA is an automatic agent that assists meeting participants, and is part of the larger CALO [2] effort to build a "Cognitive Assistant that Learns and Organizes" under DARPA's "Perceptive Assistant that Learns" (PAL) program [3]. The focus of CALO in general is "learning in the wild", or continuous improvement of the system's abilities as a result of system use. This agenda fits nicely into our goal of recognizer robustness via unsupervised adaptation. As described below, the CALO task provides meeting data that has properties that are consistent with adaptation over time, and allows us to study adaptation with varying granularity, i.e., to the general domain, to the topic, or to the speakers in question.

2. RECOGNITION SYSTEM

The baseline system for all our experiments is the meeting recognition system jointly developed by ICSI and SRI for the NIST RT-05S meeting recognition evaluation [4]. This system and its variants have shown state-of-the-art performance in the 2004, 2005, and 2006 NIST evaluations.

The recognizer performs a total of 7 decoding passes with alternating acoustic front-ends: one based on MFCCs augmented with discriminatively estimated multilayer-perceptron (MLP) features, and one based on PLP features. Acoustic models are cross-adapted during recognition to output from previous recognition stages, and the output of the three final decoding steps is combined via confusion networks. The speaker-independent acoustic models were first trained on about 2300 hours of telephone conversations using the minimum phone error criterion, and then MMI-MAP-adapted to 104 hours of meeting speech from a variety of sources. The feature MLPs were also first trained on telephone speech and then adapted to meeting speech.

To limit the scope of our study, we only investigate across-meeting adaptation of the language model in this paper. (Acoustic models are adapted to speakers within meetings as described above, but not across meetings.) The recognizer uses Kneser-Ney-smoothed bigram, trigram, and 4-gram LMs at various stages of decoding. The baseline LMs are constructed by static interpolation of models from different sources, including (non-CALO) meeting transcripts, topical telephone conversations, web data, and news; details can be found in [5]. When adapting the LMs using the strategies described below, all versions of the LM used in the recognition system (bigram, trigram, 4-gram) were adapted similarly.

3. PRIOR WORK

Adaptation methods were first proposed and are now extensively used for acoustic models. Two very popular approaches are max-

imum likelihood linear regression (MLLR) [6] and maximum a-posteriori (MAP) adaptation [7]. In MAP adaptation, a new model $\hat{\Phi}$ is computed such that

$$\hat{\Phi} = \arg\max_{\Phi}[f(W|\Phi) \cdot g(\Phi)]$$

where $f(W|\Phi)$ is the discrete density function of W and $g(\Phi)$ is the prior distribution, which is typically modeled using a Dirichlet density [7].

For LM adaptation two popular approaches are model interpolation and count mixing. In model interpolation, an out-of-domain model θ_{OOD} is interpolated with an in-domain model θ_{ID} to form an adapted model $\hat{\theta}$:

$$P_{\hat{\theta}}(w_i|h_i;\gamma) = \gamma P_{\theta_{OOD}}(w_i|h_i) + (1-\gamma)P_{\theta_{ID}}(w_i|h_i) \quad (1)$$

where $P_\theta(w_i|h_i)$ is the probability of the current word w_i given the history of $n-1$ words, h_i, in an n-gram LM θ. γ is a weight controlling the influence of the out-of-domain data on the final model and is usually optimized on a development set.

Another approach to LM adaptation is count mixing, where the n-gram counts from all sources are summed, often after applying a source-specific weights. Bacchiani and Roark have shown that both approaches are actually equivalent, and are furthermore equivalent to MAP adaptation with a different parameterization of the prior distribution [8]. They reported positive results using unsupervised LM adaptation in a voicemail recognition system.

Kneser *et al.* have proposed using dynamic marginals for model adaptation [9]. The idea is to adjust the n-gram weights so that the unigram marginals of the adapted n-gram matches the unigram distribution of the adaptation data.

Gretter and Riccardi have exploited word confidences obtained from word confusion networks during unsupervised LM adaptation [10]. Hakkani-Tür *et al.* have employed unsupervised LM adaptation for new call center spoken dialog applications [11]. One difference in their approach is that they effectively set $\gamma = 0$ so as to make the new model small enough for a sub-real-time ASR system. Previous work on conversational telephone speech recognition showed small gains with unsupervised LM adaptation to Switchboard recognition output even at fairly high error rates [12].

A research area related to unsupervised LM adaptation deals with strategies for selecting adaptation data. Some notable studies addressing this issue include [13, 14]. An extensive survey of LM adaptation research can be found in [15].

4. ADAPTATION APPROACH

In CALO-MA, our goal is to improve ASR performance using audio data from previous meetings. Since no manual transcriptions are available, we use the automatic transcriptions of these meetings to build the in-domain LM, and adapt the generic model using an interpolation approach as in (1).

Recognition is offline, allowing us to estimate the optimal adaptation weight γ on a held-out set. The only problematic issue is that, unlike in most prior work, no manual transcriptions are available for estimating γ. Instead, we again use the ASR output for the held-out set. This is similar to previous work by Niesler and Willett [16]. Below we report experiments showing that using errorful transcripts for estimating γ carries only a negligible penalty compared to using manual transcripts.

Table 1. Statistics of meeting sequences used in the experiments.

Sequence	# words	# speakers	# meetings
1	4895	4	5
2	3970	3	5
3	5318	4	3
4	1427	3	2
5	1653	3	5
6	3927	4	5
7	5948	4	5
8	4998	4	5

The estimation of γ is carried out by maximizing the log probability of the held-out data according to the LM:

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma} LP_\gamma = \operatorname*{argmin}_{\gamma} \sum_i \log P(w_i|h_i;\gamma) \quad (2)$$

where w_i is the ith word in the held-out set and h_i its n-gram history, $i = 1, \ldots, k$, and $P(\cdot)$ is computed as in (1). This is equivalent to minimizing the LM perplexity PP_γ on the held-out set, since

$$PP_\gamma = e^{-\frac{1}{k}LP_\gamma} \quad (3)$$

where k is the number of tokens in the text set. The optimization of γ is carried out by an expectation-maximization algorithm, and typically converges in a dozen or so iterations.

The steps involved in our adaptation experiments are:

1. Using the out-of-domain generic model θ_{OOD} to recognize the in-domain data to obtain the automatic transcripts $\hat{W}_{adapt}$, and the held-out set to obtain $\hat{W}_{HO}$.
2. Using $\hat{W}_{adapt}$ build a model θ_{ID}.
3. Interpolate θ_{OOD} and θ_{ID} to obtain $\hat{\theta}$. Use the interpolation weight γ that maximizes the log probability of $\hat{W}_{HO}$.

As we receive more data to recognize, $\hat{W}_{adapt}$ is enlarged and the above steps are repeated. This can be done using the original model as θ_{OOD}, or the model resulting from the most recent adaptation step. In the next section, we compare the effectiveness of these two approaches.

All LMs are built using modified Kneser-Ney smoothing [17, 18] using the SRILM [19] toolkit. The adapted LMs were constructed using the static n-gram interpolation functionality in SRILM, merging the n-grams of the baseline and the in-domain LMs into a single new backoff LM, and assigning them interpolated conditional probabilities.

5. EXPERIMENTS AND RESULTS

5.1. Meetings Data

For the CALO-MA project, SRI collected 8 sequences of meetings, each with up to 5 meetings, totaling 35 meetings with 32,136 words. There are 10 speakers in total, with the same speakers (with some exceptions) occurring throughout a meeting sequence, but also re-occurring across sequences. Each sequence contains meetings on a coherent topic (such as hiring new staff). Some statistics describing the meeting sequences are given in Table 1.

In the remainder of this section we report on three sets of LM adaptation experiments: *across-sequence* (where data from the previous sequences are used for adaptation), *within-sequence* (where data from meetings in a sequence are used for adaptation), and *within-speaker* adaptation.

Table 2. Perplexities using unsupervised and supervised adaptation methods and interpolation weight training.

LM Training	Weight Estimation	Perplexity
Baseline	n/a	101.7
Unsupervised	Unsupervised	93.1
Unsupervised	Supervised	91.9
Supervised	Unsupervised	85.7
Supervised	Supervised	85.6

Table 3. Across-meeting-sequence adaptation experiments using unsupervised (with only ASR output) and supervised (with manual transcriptions) LM adaptation.

Model	WER
Baseline	16.2%
Unsupervised	15.3%
Supervised	14.4%
Unsupervised4	15.4%
Supervised4	14.0%
Unsupervised+4	15.5%
Supervised+4	14.3%

5.2. Across-Sequence Adaptation

We performed across-sequence adaptation using sequences 1 and 2 for training, sequences 5 and 6 for tuning, and sequences 7 and 8 for testing. (Sequences 3 and 4 were set aside for a follow-on experiment, described below.) The baseline performance is obtained using the generic LM. To evaluate the effect of the errors introduced by ASR output, we also ran a control experiment in which the LM was adapted using the manual transcriptions, instead of the ASR output. However, to keep the results as comparable as possible, no new words were added to the recognizer vocabulary.[1]

Table 2 presents perplexities of all models on the manual transcriptions of the test set. Where applicable, we compared the adapted models with interpolation weights estimated using both manual (supervised) and automatic (unsupervised) transcriptions of the held-out set. The results show that, for supervised adaptation, this distinction did not matter at all. For unsupervised learning, perplexity increased only slightly when estimating the interpolation weight on ASR transcripts. Both results confirm that adaptation weight estimation is robust to ASR errors.

Table 3 presents our ASR results. Unsupervised adaptation reduced the word error rate (WER) significantly[2] from 16.2% to 15.3%. With supervised adaptation, WER dropped to 14.4%. The relative WER reductions are 5.5% and 11.1%, respectively.

Next, we used sequences 3 and 4 in two ways, either by combining them with sequences 1 and 2 ("Supervised4" and "Unsupervised4"), or to test incremental adaptation, that is, to adapt the model that was already adapted to sequences 1 and 2 (giving "Supervised+4" and "Unsupervised+4"). While the performance does not change significantly for either unsupervised or supervised adaptation, pooling all the data performed the best for supervised adaptation.

[1] Unsupervised adaptation by definition does not modify the vocabulary since the adaptation data can only contain in-vocabulary words.

[2] Using a matched-pair sign test with $p < 0.0005$.

Table 4. Within-meeting-sequence adaptation experiments using supervised and unsupervised LM adaptation.

Model	WER
Baseline	28.6%
Unsupervised	27.1%
Supervised	25.2%

Table 5. Speaker adaptation experiments using supervised and unsupervised LM adaptation.

Model	WER
Baseline	15.0%
Unsupervised	13.9%
Supervised	12.2%

5.3. Within-Sequence Adaptation

CALO-MA meetings are set up to discuss a given topic with a specific agenda and action items, although several sequences are similar in terms of their topics. The meetings are therefore modeled on the kinds of meetings that would take place in real-life organizations, with threads of recurring topics running through several meetings. This motivates the second set of adaptation experiments, where we used only the first three meetings in a given sequence for adaptation, the fourth meeting for weight estimation, and the last meeting for testing. Each test meeting is then recognized using its own specialized LM. Sequences 3 and 4 were excluded from this experiment since they lacked the requisite number of meetings.

Table 4 presents the WERs aggregated over all sequences. Similar to the previous experiment, we achieved significant WER reductions. In relative terms, the error reductions were almost identical to the across-meeting results: 5.2% using unsupervised adaptation and 11.9% using supervised adaptation. (The absolute error rates differ due to the different choice of test set.) We also noted that ASR accuracy improved for each individual meeting sequence, not just in overall terms.

5.4. Within-Speaker Adaptation

The next set of experiments explores adaptation to the data of each individual speaker across meetings, for those speakers who participated in more than one meeting. Similar to the within-sequence adaptation experiments, instead of using a separate adapted LM for each sequence, we used a separate adapted LM for each speaker. In this mode, we give the LM an opportunity to capture speakers' idiosyncratic speaking styles, rather than (just) domain or topic characteristics. In this mode, it is also possible that the LM adapts to subjects that are typically covered by a speaker's meeting contributions (such as when a speaker is an expert for a particular topic). Note that, unlike the way offline acoustic speaker adaptation is usually carried out, the adaptation data consists only of data that (temporally) precedes the data being recognized, and excludes the test data.

Table 5 presents our results, in terms of aggregate WERs across all speakers. Word error rate again dropped significantly: 7.3% relative using unsupervised adaptation and 18.7% relative using supervised adaptation.

Table 6. Comparing different adaptation methods for supervised and unsupervised LM adaptation. "# Words"indicate the amount of data used for adaptation.

	Within-Sequence	Across-Sequence	Speaker
# Words	2,824	15,610	9,640
Baseline	15.8%	15.8%	15.8%
Unsupervised	14.3%	14.3%	14.3%
Supervised	12.1%	13.4%	14.3%

5.5. Comparing Adaptation Methods

The previous experiments raise the question which adaptation strategy might be optimal for the given task. While one could explore combined strategies, here we simply performed a side-by-side comparison using a common test set, consisting of the last meeting of Sequence 8. This meeting comprises 841 words. Table 6 presents the amount of data used for adaptation and the results using within-sequence, across-sequence, and within-speaker adaptation, in both supervised and unsupervised modes.

Similar to previous experiments, the WER was reduced by 8.3%-9.6% relative using unsupervised adaptation, and 9.6%-22.4% relative using supervised adaptation. When manually transcribed data is available, within-sequence adaptation outperforms the others, even though it uses the smallest amount of data. Across-sequence adaptation performed the worst, although it exploited the largest amount of data for adaptation. This seems to indicate that at least in this domain, specificity of the adaptation data is more important than quantity. However, when we compare unsupervised adaptation results, we see a somewhat different pattern, without significant differences between the three methods.

6. CONCLUSIONS

We have presented experimental results using unsupervised adaptation of LMs for the recognition of agenda-driven meetings with temporal structure across meetings. Different ways to select adaptation data were investigated, including across-meetings, within-sequence, and within-speaker. We obtained significant error rate reductions of between 5% and 9%, using LMs adapted by interpolation of the generic LM with LMs constructed from automatic ASR output. All data selection modes gave similar results (although larger test corpora might yet show significant differences between these strategies). The improvements seen are generally about half of what could be achieved with supervised adaptation (to human transcripts) using the same data.

In future work, we plan to investigate the use of word confidence estimates and lattice hypotheses to gain more leverage from errorful ASR hypotheses in adaptation. Furthermore, LMs could be adapted to both prior ASR output and written documents (such as agendas, emails, or Web data) related to a target meeting.

7. ACKNOWLEDGMENTS

We thank Dilek Hakkani-Tür and Elizabeth Shriberg for many helpful discussions. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

8. REFERENCES

[1] Jonathan G. Fiscus, Nicolas Radde, John S. Garofolo, Audrey Le, Jerome Ajot, and Christophe Laprun, "The Rich Transcription 2005 Spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, Steve Renals and Samy Bengio, Eds. 2006, vol. 3869 of *Lecture Notes in Computer Science*, pp. 369–389, Springer.

[2] "SRI Cognitive Assistant that Learns and Organizes (CALO) Project," http://www.ai.sri.com/project/CALO.

[3] "DARPA Perceptive Assistant that Learns (PAL) Program," http://www.darpa.mil/ipto/programs/pal/index.htm.

[4] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Proceedings of the NIST Meeting Recognition Workshop*, Edinburgh, UK, 2005.

[5] Özgür Çetin and Andreas Stolcke, "Language modeling in the ICSI-SRI Spring 2005 meeting speech recognition evaluation system," Tech. Rep. TR-05-06, International Computer Science Institute, Berkeley, CA, 2005.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[7] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[8] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[9] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals," in *Proceedings of the EUROSPEECH*, Rhodes, Greece, September 1997.

[10] R. Gretter and G. Riccardi, "On-line learning of language models with word error probability distributions," in *Proceedings of the ICASSP*, Salt Lake City, Utah, May 2001.

[11] D. Hakkani-Tür, G. Tur, M. Rahim, and G. Riccardi, "Unsupervised and active learning in automatic speech recognition for call classification," in *Proceedings of the ICASSP*, Motreal, Canada, May 2004.

[12] A. Stolcke, "Error modeling and unsupervised language modeling," in *Proceedings of the NIST LVCSR Workshop*, Linthicum, MD, 2001.

[13] L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[14] H. Nanjo and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," in *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.

[15] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," *Speech Communication Special Issue on Adaptation Methods for Speech Recognition*, vol. 42, pp. 93–108, 2004.

[16] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[17] Reinhard Kneser and Hermann Ney, "Improved clustering techniques for class-based statistical language modeling," in *Proc. EUROSPEECH*, Berlin, Sept. 1993, vol. 2, pp. 973–976.

[18] Stanley F. Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," Tech. Rep. TR-10-98, Computer Science Group, Harvard University, Aug. 1998.

[19] Andreas Stolcke, "SRILM—an extensible language modeling toolkit," in *Proc. ICSLP*, John H. L. Hansen and Bryan Pellom, Eds., Denver, Sept. 2002, vol. 2, pp. 901–904.