EXPLOITING USER FEEDBACK FOR LANGUAGE MODEL ADAPTATION IN MEETING RECOGNITION

Dimitra Vergyri *Andreas Stolcke* *Gokhan Tur*

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

ABSTRACT

We investigate language model (LM) adaptation in a meeting recognition application, where the LM is adapted based on recognition output from relevant prior meetings and partial manual corrections. Unlike previous work, which has considered either completely unsupervised or supervised adaptation, we investigate a scenario where a human (e.g., a meeting participant) can correct some of the recognition mistakes. We find that recognition accuracy using the adapted LM can be enhanced substantially by partial correction. In particular, if all content words (about half of all recognition errors) are corrected, recognition improves to the same accuracy as if completely error-free (manually created) transcriptions had been used for adaptation. We also compare and combine a variety of adaptation methods, including linear interpolation, unigram marginal adaptation, and a discriminative method based on "positive" and "negative" N-grams.

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

1. INTRODUCTION

A continuing challenge for speech recognition is the domain and style mismatch between language model (LM) training data and target application. An attractive potential solution is unsupervised adaptation, that is, the notion that relevant adaptation data can be automatically transcribed and, given low-enough error rates, serve as additional training data for the LM. Past work has shown partial success with this approach, albeit with a substantial degradation compared to supervised adaptation, i.e., the scenario where human transcripts of the adaptation data are available [1, 2].

We investigate a hybrid adaptation method where humans generate partially corrected transcripts of relevant data. As in [2], our application domain is meeting recognition where a series of multiparty conversations related by topic and/or style is processed, and each meeting becomes potential adaptation data for subsequent meetings. In such a scenario it is not unreasonable to present meeting participants with automatic transcripts of their own contributions, and allow them to correct recognition errors that are particularly egregious. In fact, meeting participants will often want to so do out of self-interest, to "set the the record" straight. We are in the process of implementing such a user-feedback facility as part of the CALO Meeting Assistant System [3].

We investigate the effectiveness of user correction for LM adaptation purposes, by simulating partial transcript corrections in our manually transcribed data. We also compare and combine a variety of adaptation methods, including linear interpolation, unigram marginal adaptation, and a nonstandard method based on discriminative reranking of recognition hypotheses with "positive" and "negative" N-gram LMs.

In the literature, user feedback is typically exploited in dictation systems. Yu *et al.* have proposed a speech-based correction method in a dictation system, where the focus is learning new words using a phoneme recognizer [4]. The language model is updated using a method similar to cache language modeling [5] only for the rest of the session. Ogata and Goto, on the other hand, have proposed using confusion networks to ease the user correction task [6]. We are not aware of a work which changes the language model based on user corrections for the following sessions.

In the next section, we will briefly describe the Decipher automatic speech recognition system used in this study. Then Section 3 presents the language model adaptation methods employed to exploit the user feedback. Section 4 shows the performance figures obtained from the experiments simulating user feedback.

2. RECOGNITION SYSTEM

The baseline system for all our experiments is the meeting recognition system jointly developed by SRI and ICSI for the NIST RT-05S meeting recognition evaluation [7]. This system and its variants have shown state-of-the-art performance in the 2004, 2005, 2006, and 2007 NIST evaluations.

The recognizer performs a total of seven decoding passes with alternating acoustic front-ends: one based on Mel frequency cepstral coefficients (MFCCs) augmented with discriminatively estimated multilayer-perceptron (MLP) features, and one based on perceptual linear prediction (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 adapted to 104 hours of meeting speech from a variety of sources. The feature MLPs were 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, i.e., the LM is adapted before beginning recognition of a meeting, and held fixed throughout. (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 [8].

Meeting 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

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

3. ADAPTATION METHODS

Our experiments employ a variety of LM adaptation techniques, alone or in combination.

3.1. Linear Interpolation

In this computationally simple and time-honored approach (going back at least to [9]), a separate LM p_{adapt} is estimated from the transcripts (automatic or human generated) of the adaptation data. The adapted LM is then obtained by forming the linear interpolation of the baseline LM p_{base} with p_{adapt}:

$$p(w|h) = \lambda p_{base}(w|h) + (1-\lambda) p_{adapt}(w|h) \quad (1)$$

The adaptation weight λ is optimized on a held-out set. In our implementation, rather than optimizing word error directly, an expectation maximization algorithm is used to minimize the perplexity of the tuning set, as a function of λ. As we found in prior work [2], no significant degradation is incurred by estimating λ on error transcripts rather than correct transcripts.

If the interpolated LM components are both N-gram models one can create a single new merged LM that incorporates the interpolated probabilities [10]. Therefore, it is straightforward to adapt all LMs used by the recognition system (including the one used in decoding).

3.2. Unigram Marginal Adaptation

In this approach the baseline LM is modified to change the unigram marginal probabilities to match the adaptation data, but higher-order N-gram probabilities are changed as little as possible (in the relative entropy sense). This approach was proposed by [11]; a fast, approximate version was suggested by [12] in which the adapted LM is

$$p(w|h) = \frac{p_{base}(w|h) \left(\frac{p_{adapt}(w)}{p_{base}(w)} \right)^{\beta}}{Z(h)} \quad (2)$$

where $Z(h)$ is a normalizing term that ensures the probabilities for a given history h sum to unity, i.e.,

$$Z(h) = \sum_{w} p_{base}(w|h) \left(\frac{p_{adapt}(w)}{p_{base}(w)} \right)^{\beta} \quad (3)$$

β is an attenuation parameter controlling how far the adapted estimates deviate from the baseline.

The rationale behind unigram marginal adaptation is that while higher-order N-gram estimates based on small and errorful adaptation data are bound to be noisy, unigram probabilities should be relatively reliable.

Due to the normalization term (3), marginal adaptation is computationally expensive. We therefore used it only in the N-best rescoring stages of our system. An implementation [10] that caches $Z(h)$ for reuse can be very efficient for a typical rescoring application since the same history occurs many times.

3.3. Positive / Negative LM Rescoring

The third adaptation method is meant to be discriminative in that it actively penalizes N-grams that were recognized incorrectly in the adaptation data. This approach assumes that at least some manually corrected recognition output is available.

To this end, we compare the automatic hypothesis to the corrected ones and extra two sets of N-grams: those that occur in the corrected transcripts in regions of recognition errors (the "positive" N-grams), and those N-grams containing the recognition errors themselves (the "negative" N-grams). From these N-gram counts, corresponding "positive" and "negative" LMs are estimated, and N-best are rescored with each LM. The resulting additional scores are then combined in a log-linear, weighted fashion with the standard scores for acoustic and baseline language models. The combination weights are then optimized discriminatively to minimize word errors on the held-out tuning set. The "negative" LM scores typically receive a negative weight in this process, i.e., this results in a penalty to hypotheses containing N-grams found in the incorrect adaptation hypotheses. The method is similar to the "anti-language model" proposed in [13].

Also note that this approach resembles the minimum classification (or word) error (MCE or MWE) based discriminative language model training approaches [14]. In MCE-based language modeling the probability of the N-grams in the correct (erroneous) string are increased (decreased) a little. More formally, a misclassification function is defined to compute the difference between the language model scores (log probabilities), G_{LM} of the correct string, W_0, and the best hypotheses, W_1 for a given portion of speech, X_i:

$$d_{LM}(X_i) = -g_{LM}(X_i, W_0) + g_{LM}(X_i, W_1)$$

Then the goal is to minimize this misclassification function. A loss function is defined as:

$$\frac{1}{1 + exp(-\gamma d_{LM}(X_i) + \theta)}$$

Then the generalized probabilistic descent (GPD) algorithm is employed to update each N-gram score in an iterative fashion. In that respect, our approach can be considered as a simplified case of a MCE-based approach since we learn a single weight for all the adjusted N-grams and we do not iterate.

4. EXPERIMENTS

4.1. Meeting Data

For the CALO-MA project [3], SRI collected eight sequences of meetings, each with as many as five meetings, totaling 35 meetings with 32,136 transcribed 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.

We performed across-sequence adaptation using meeting sequences 1 through 4 as adaptation data, sequences 5 and 6 for tuning

Line	Model	WER (%)
1	Unadapted	16.1
2	Adapted: interp(unsup)	15.4
3	Adapted: interp(unsup) + marg(unsup)	15.4
4	Adapted: interp(unsup) + pos/neg(partial)	15.0
5	Adapted: interp(unsup) + marg(unsup) + pos/neg(partial)	15.2
6	Adapted: marg(unsup) + pos/neg(partial)	15.0
7	Adapted: interp(unsup) + pos/neg(sup)	14.9
8	Adapted: interp(sup)	14.0

Table 2. Meeting recognition results on the test set using unadapted and various adapted language models for rescoring N-best hypotheses.

Line	Model	approx. % of content words corrected	% all errors corrected	WER (%)	f-WER (%) (rel. improv.)		c-WER (%) (rel. improv.)	
0	Unadapted	0 %	0 %	16.1	19.4		12.0	
1	Adapted: interp(unsup)	0 %	0 %	15.4	18.5	(-4.6%)	11.3	(-6%)
2	Adapted: interp(partial)	25 %	14 %	15.0	18.3	(-5.7%)	10.8	(-10%)
3	Adapted: interp(partial)	50 %	29 %	14.7	18.0	(-7.2%)	10.4	(-13%)
4	Adapted: interp(partial)	100 %	55 %	14.0	17.6	(-9.3%)	9.4	(-22%)
5	Adapted: interp(sup)	100 %	100 %	14.0	17.4	(-10.3%)	9.5	(-21%)

Table 3. Meeting recognition WER results with varying percentages of content-word errors corrected in adaptation data. Results are also reported separately for function words (f-WER) and content words (c-WER) along with the relative improvement for these word categories.

and estimation of free parameters, and sequences 7 and 8 for testing. The baseline performance is obtained using the generic LM. Contrasting experiments employ unsupervised and partially supervised adaptation, using methods described in Section 3. As an upper bound on adaptation performance, we also ran supervised LM adaptation using the manual transcriptions of the adaptation meetings. Note that supervised and partially supervised adaptation also implies adding previously unseen words that occur in the adaptation hypotheses to the recognizer vocabulary.

In order to simulate user corrections, we assumed that users do not correct function (or stop) words such as *so* or *the* unless they are part of a larger sequence. Then we simulated user feedback at various correction levels. For example the word *the* is also restored in the word sequence *join the CALO project* erroneously recognized as *joined kayla project*. We do not present the recognition performance figures for the corrected meetings since we assume that this is not a functionality of user interest. Instead we only provide performance figures for the following meetings (which can be of completely different subjects).

4.2. Results and Discussion

Table 2 summarizes the test set word error rates for the unadapted (baseline) and various adapted language models. For the adapted LMs we encode the combination of methods used as follows. "interp" refers to linear interpolation, "marg" means unigram marginal adaptation, and "pos/neg" refers to rescoring with positive/negative LMs. The hypotheses used in each case are given in parentheses after the method: "unsup" for raw recognizer hypotheses, "sup" for manual (fully corrected) transcripts, and "partial" for partially corrected transcripts. Note that, in line 6, the LM adapted by interpolation was not used in N-best rescoring, but was used in generating the N-best lists, since the other adaptation methods can be used only in rescoring.

Marginal adaptation does not seem to give an added benefit over linear interpolation (lines 2 and 3, as well as 4 and 6) on our data; we therefore did not explore other combinations involving marginals adaptation since it is computationally expensive. Also, the fact that adding the marginal-adapted LM to the other two LMs actually degrades performance (line 5) indicates that a larger tuning set might be needed to robustly estimate the required rescoring weights.

The relative error reduction over the unadapted baseline from supervised adaptation is 13% (lines 1 and 8), and unsupervised adaptation (line 4) recovers about one third of that gain (4.3% relative). However, if we employ partial correction, the gain improves to 6.8% relative (line 4).

Finally, it seems that rescoring with fully corrected positive/negative LMs (line 7) gives only a small (0.1% absolute) gain over the partially corrected version (line 4).

We now turn to a system in which all LMs (both those used in decoding and those used in rescoring) had been adapted to partially corrected hypotheses obtained from the training data. To simulate plausible user behavior, we corrected only recognition errors in regions involving content words, and varied the percentage of corrected errors from about 14% (with a quarter of all content-word errors corrected) to 55% (with all content-word errors corrected). The results are summarized in Table 3.

As expected, performance improves steadily as more and more content word errors in the adaptation data are corrected. With half of all content-word errors (29% of all errors) corrected, the gain over the unsupervised approach is 0.7% absolute, or half of the gain obtained with fully supervised adaptation.

Maybe surprisingly, however, correcting all content-word error regions (55% of all errors) gives the same performance as correcting all errors (i.e., fully supervised adaptation). A possible explanation is that the LM is already well-trained for non-content words, and therefore can only be improved in the modeling of content words. This can further be demonstrated by looking at the f-WER and c-WER that are computed over function and content words respectively. We see that for function words with supervised adaptation we get only about 10% of the errors corrected, while we get half of this improvement just with unsupervised adaptation. On the other hand,

for content words, with supervised adaptation we fix a bit more than 20% of the errors, while only a quarter of these (6% of the content word errors) were fixed with unsupervised adaptation. This result shows that the user feedback for LM adaptation can benefit mostly the content words, and that it is not necessary to fix all word errors in order to get the full benefit of supervised adaptation.

In a realistic scenario where the user corrects certain recognizer errors such as names or abbreviations, meeting after meeting, we expect to see significant performance improvement in recognizing them using the partially supervised LM adaptation approach.

5. CONCLUSIONS AND FUTURE WORK

We have explored language model adaptation in a meeting recognition scenario, specifically investigating an adaptation mode where a human user or annotator partially corrects transcriptions of the adaptation data. The proposed method is compared to completely unsupervised and completely supervised adaptation. We found that by correcting half the recognition errors and focusing on errors involving content words, the same error reduction was achieved as with completely supervised adaptation. That improvement was a 13% relative WER reduction over the unadapted system, and a 9% relative WER over completely unsupervised adaptation. A "lazy" correction of only a fraction the content word errors in the adaptation transcripts yielded correspondingly worse performance, with unsupervised adaptation as a lower bound. Taken together, these results suggest that allowing meeting participants to give the recognizer feedback by correcting some of the important misrecognitions in their own data would an effective way to improve system performance over time, giving substantial gain over simple unsupervised adaptation.

These results were obtained by applying linear interpolation to all the LMs used by our multi-pass recognition system. Additional adaptation methods, like unigram marginal adaptation and discriminative modeling of correct/incorrect N-grams, did not give additional wins.

Future work will include a more exhaustive investigation of the other adaptation methods on partially corrected hypotheses, such as MCE, and a thorough error analysis. In particular, we would like to see if the error reductions on the test data also affect mostly content words, in accordance with how the partial corrections were selected.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract FA8750-07-D-0185 / DO 0004. 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 DARPA or Air Force Research Laboratory.

7. REFERENCES

[1] M. Bacchiani and B. Roark, "Unsupervised language model adaptation", *in Proc. ICASSP*, vol. 1, pp. 224–227, Hong Kong, Apr. 2003.

[2] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition", *in Proc. ICASSP*, vol. 4, pp. 173–176, Honolulu, Apr. 2007.

[3] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting speech recognition and understanding system", *in Proc. IEEE Spoken Language Technology Workshop*, Goa, Dec. 2008.

[4] D. Yu, M. Hwang, P. Mau, A. Acero, and L. Deng, "Unsupervised learning from users' error correction in speech dictation", *in Proceedings of the ICSLP*, Jeju-Island, Korea, October 2004.

[5] R. Kuhn and R. de Mori, "A cache-base natural language model for speech recognition", *IEEE PAMI*, vol. 12, pp. 570–583, June 1990.

[6] J. Ogata and M. Goto, "Speech repair: Quick error correction just by using selection operation for speech input interfaces", *in Proceedings of the Interspeech*, Lisbon, Portugal, September 2005.

[7] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, 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 Rich Transcription 2005 Spring Meeting Recognition Evaluation*, pp. 39–50, Edinburgh, July 2005. National Institute of Standards and Technology.

[8] Ö. Çetin and A. Stolcke, "Language modeling in the ICSI-SRI Spring 2005 meeting speech recognition evaluation system", Technical Report TR-05-06, International Computer Science Institute, Berkeley, CA, 2005.

[9] F. H. Liu, M. D. Monkowski, M. Novak, M. Padmanabhan, M. A. Picheny, and P. S. Rao, "IBM Switchboard progress and evaluation site report", *in LVCSR Workshop*, Gaithersburg, MD, Apr. 1995. National Institute of Standards and Technology.

[10] A. Stolcke, "SRILM—an extensible language modeling toolkit", in J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, vol. 2, pp. 901–904, Denver, Sep. 2002.

[11] P. S. Rao, M. D. Monkowski, and S. Roukos, "Language model adaptation via minimum discriminant information", *in Proc. ICASSP*, pp. 161–164, Detroit, May 1995.

[12] R. Kneser, J. Peters, and D. Klakow, "Language model adaptation using dynamic marginals", in G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 4, pp. 1971–1974, Rhodes, Greece, Sep. 1997.

[13] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system", *in Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[14] H.-K. J. Kuo, E. Fosler-Lussier, H. Jiang, and C.-H. Lee, "Discriminative training of language models for speech recognition", *in Proc. ICASSP*, vol. 1, pp. 325–328, Orlando, FL, May 2002.