Implicitly Supervised Language Model Adaptation for Meeting Transcription

David Huggins-Daines
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dhuggins@cs.cmu.edu

Alexander I. Rudnicky
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
air@cs.cmu.edu

Abstract

We describe the use of meeting metadata, acquired using a computerized meeting organization and note-taking system, to improve automatic transcription of meetings. By applying a two-step language model adaptation process based on notes and agenda items, we were able to reduce perplexity by 9% and word error rate by 4% relative on a set of ten meetings recorded in-house. This approach can be used to leverage other types of metadata.

1 Introduction

Automatic transcription of multi-party conversations such as meetings is one of the most difficult tasks in automatic speech recognition. In (Morgan et al., 2003) it is described as an "ASR-complete" problem, one that presents unique challenges for every component of a speech recognition system.

Though much of the literature on meeting transcription has focused on the unique acoustic modeling and segmentation problems incurred by meeting transcription, language modeling for meetings is an interesting problem as well. Though meeting speech is spontaneous in nature, the vocabulary and phrasing in meetings can be very specialized and often highly technical. Speaking style can vary greatly between speakers, and the discourse structure of multi-party interaction gives rise to cross-speaker effects that are difficult to model with standard N-gram models (Ji and Bilmes, 2004).

Speech in meetings has one crucial advantage over many other transcription tasks, namely that it does not occur in isolation. Meetings are scheduled and discussed in advance, often via e-mail. People take notes and create agendas for meetings, and often read directly from electronic presentation materials. The structure of meetings can be exploited - topics can be segmented both temporally and across speakers, and these shifting topics can be modeled as sub-languages.

We examine the effect of leveraging one particular type of external information, namely the written agendas and meeting minutes, and we demonstrate that, by using off-line language model adaptation techniques, these can significantly ($p < 0.01$) improve language modeling and speech recognition accuracy. The language in the notes and agendas is very similar to that used by the speakers, hence we consider this to be a form of semi-supervised or *implicitly supervised* adaptation.

2 Corpus

The SmartNotes system, described in (Banerjee and Rudnicky, 2007) is a collaborative platform for meeting organization, recording, and note-taking. As part of our research into meeting segmentation and recognition, we have collected a series of 10 unscripted meetings using SmartNotes. These meetings themselves are approximately 30 minutes in length (ranging from 1745 to 7208 words) with three regular participants, and consist of discussions and reporting on our ongoing research. The meetings are structured around the agendas and action items constructed through the SmartNotes interface. The

Proceedings of NAACL HLT 2007, Companion Volume, pages 73–76,
Rochester, NY, April 2007.

agenda itself is largely constant from meeting to meeting, while each meeting typically reviews discusses the previous week's action items. Each participant is equipped with a laptop computer and an individual headset microphone.

Each meeting was manually transcribed and segmented for training and testing purposes. The transcription includes speaker identification and timing information. As part of the meeting, participants are encouraged to take notes and define action items. These are automatically collected on a server along with timestamp information. In (Banerjee and Rudnicky, 2007), it was shown that timestamped text of this kind is useful for topic segmentation of meetings. In this work, we have not attempted to take advantage of the timing information, nor have we attempted to perform any topic segmentation. Given the small quantity of text available from the notes, we feel that the type of static language model adaptation presented here is most feasible when done at the entire meeting level. A cache language model (Kuhn and Mori, 1990) may be able to capture the (informally attested) locality effects between notes and speech.

Since the notes are naturalistic text, often containing shorthand, abbreviations, numbers, punctuation, and so forth, we preprocess them by running them through the text-normalization component of the Festival[1] speech synthesis system and extracting the resulting string of individual words. This yielded an average of 252 words of adaptation data for each of the 10 meetings.

3 System Description

Unless otherwise noted, all language models evaluated here are trigram models using Katz smoothing (Katz, 1987) and Good-Turing discounting. Linear interpolation of multiple source models was performed by maximizing the likelihood over a held-out set of adaptation data.

For automatic transcription, our acoustic models consist of 5000 tied triphone states (senones), each using a 64-component Gaussian mixture model with diagonal covariance matrices. The input features consist of 13-dimensional MFCC features, delta, and delta-delta coefficients. These models are trained on approximately 370 hours of speech data, consisting of the ICSI meeting corpus (Morgan et al., 2003), the HUB-4 Broadcast News corpus, the NIST pilot meeting corpus, the WSJ CSR-0 and CSR-1 corpora,[2] the CMU Arctic TTS corpora (Kominek and Black, 2004), and a corpus of 32 hours of meetings previously recorded by our group in 2004 and 2005.

Corpus	# Words	Perplexity
Fisher English	19902585	178.41
Switchboard-I	2781951	215.52
ICSI (75 Meetings)	710115	134.94
Regular Meetings	266043	111.76
Switchboard Cellular	253977	280.81
CallHome English	211377	272.19
NIST Meetings	136932	199.40
CMU (ISL Meetings)	107235	292.86
Scenario Meetings	36694	306.43

Table 1: Source Corpora for Language Model

Our baseline language model is based on a linear interpolation of source language models built from conversational and meeting speech corpora, using a held-out set of previously recorded "scenario" meetings. These meetings are unscripted, but have a fixed topic and structure, which is a fictitious scenario involving the hiring of new researchers. The source language models contain a total of 24 million words from nine different corpora, as detailed in Table 1. The "Regular Meetings" and "Scenario Meetings" were collected in-house and consist of the same 32 hours of meetings mentioned above, along with the remainder of the scenario meetings. We used a vocabulary of 20795 words, consisting of all words from the locally recorded, ICSI, and NIST meetings, combined with the Switchboard-I vocabulary (with the exception of words occurring less than 3 times). The Switchboard and Fisher models were pruned by dropping singleton trigrams.

4 Interpolation and Vocabulary Closure

We created one adapted language model for each meeting using a two-step process. First, the source language models were re-combined using linear interpolation to minimize perplexity on the set of notes

[1] http://www.festvox.org/

[2] All corpora are available through http://ldc.upenn.edu/

Meeting	Baseline	Interpolated	Closure
04/17	90.05	85.96	84.41
04/24	90.16	85.54	81.88
05/02	94.27	89.24	89.19
05/12	110.95	101.68	87.13
05/18	85.78	81.50	78.04
05/23	97.51	93.07	94.39
06/02	109.70	104.49	101.90
06/12	96.80	92.88	91.05
06/16	93.93	87.71	79.17
06/20	97.19	93.88	92.48
Mean	96.57	91.59 (-5.04)	87.96 (-8.7)
S.D.	8.61	7.21 (1.69)	7.40 (6.2)
p	n/a	< 0.01	< 0.01

Table 2: Adaptation Results: Perplexity

Meeting	Baseline	Interpolated	Closure
04/17	45.22	44.37	43.34
04/24	47.35	46.43	45.25
05/02	47.20	47.20	46.28
05/12	49.74	48.02	46.07
05/18	45.29	44.63	43.44
05/23	43.68	43.00	42.80
06/02	48.66	48.29	47.85
06/12	45.68	45.90	45.28
06/16	45.98	45.45	44.29
06/20	47.03	46.73	46.68
Mean	46.59	46.0 (-0.58)	45.13 (-1.46)
S.D.	1.78	1.68 (0.54)	1.64 (1.0)
p	n/a	< 0.01	< 0.01

Table 3: Adaptation Results: Word Error

for each meeting. Next, the vocabulary was expanded using the notes. In order to accomplish this, a trigram language model was trained from the notes themselves and interpolated with the output of the previous step using a small, fixed interpolation weight $\lambda = 0.1$. It should be noted that this also has the effect of slightly boosting the probabilities of the N-grams that appear in the notes. We felt this was useful because, though these probabilities are not reliably estimated, it is likely that people will use many of the same N-grams in the notes as in their meeting speech, particularly in the case of numbers and acronyms. The results of interpolation and N-gram closure are shown in Table 2 in terms of test-set perplexity, and in Table 3 in terms of word error rate. Using a paired t-test over the 10 meetings, the improvements in perplexity and accuracy are highly significant ($p < 0.01$).

5 Topic Clustering and Dimensionality Reduction

In examining the interpolation component of the adaptation method described above, we noticed that the in-house meetings and the ICSI meetings consistently took on the largest interpolation weights. This is not surprising since both of these corpora are similar to the test meetings. However, all of the source corpora cover potentially relevant topics, and by interpolating the corpora as single units, we have no way to control the weights given to individual topics within them. Also, people may use different, but related, words in writing and speaking to describe the same topic, but we are unable to capture these semantic associations between the notes and speech.

To investigate these issues, we conducted several brief experiments using a reduced training corpus consisting of 69 ICSI meetings. We converted these to a vector-space representation using *tf.idf* scores and used a *deterministic annealing* algorithm (Rose, 1998) to create hard clusters of meetings, from each of which we trained a source model for linear interpolation. We compared these clusters to random uniform partitions of the corpus. The interpolation weights were trained on the notes, and the models were tested on the meeting transcripts. Out-of-vocabulary words were not removed from the perplexity calculation. The results (mean and standard deviation over 10 meetings) are shown in Table 4. For numbers of clusters between 2 and 42, the annealing-based clusters significantly outperform the random partition. The perplexity with 42 clusters is also significantly lower ($p < 0.01$) than the perplexity (256.5 ± 21.5) obtained by training a separate source model for each meeting.

To address the second issue of vocabulary mismatches between notes and speech, we applied *probabilistic latent semantic analysis* (Hofmann, 1999) to the corpus, and used this to "expand" the vocabulary of the notes. We trained a 32-factor PLSA model on the content words (we used a simple

# of Clusters	Random	Annealed
2	546.5 ± 107.4	514.1 ± 97.9
4	462.2 ± 86.3	426.2 ± 73.9
8	397.7 ± 67.1	356.1 ± 54.9
42	281.6 ± 31.5	253.7 ± 22.9

Table 4: Topic Clustering Results: Perplexity

Meeting	Baseline	PLSA	"Boosted"
04/17	105.49	104.59	104.87
04/24	98.97	97.58	97.80
05/02	105.61	104.15	104.48
05/12	122.37	116.73	118.04
05/18	98.55	94.92	95.18
05/23	111.28	107.84	108.03
06/02	125.31	121.49	121.64
06/12	109.31	106.38	106.55
06/16	106.86	103.27	104.28
06/20	117.46	113.76	114.18
Mean	110.12	107.07	107.50
S.D.	8.64	7.84	7.93
p	n/a	< 0.01	< 0.01

Table 5: PLSA Results: Perplexity

entropy-based pruning to identify these "content words") from the ICSI meeting vocabulary. To adapt the language model, we used the "folding-in" procedure described in (Hofmann, 1999), running an iteration of EM over the notes to obtain an adapted unigram distribution. We then simply updated the unigram probabilities in the language model with these new values and renormalized. While the results, shown in Table 5, show a statistically significant improvement in perplexity, this adaptation method is is problematic, as it increases the probability mass given to all the words in the PLSA model. In subsequent results, also shown in Table 5, we found that simply extracting these words from the original unigram distribution and boosting their probabilities by the equivalent amount also reduces perplexity by nearly as much (though the difference from the PLSA model is statistically significant, $p = 0.004$).

6 Conclusions

We have shown that notes collected automatically from participants in a structured meeting situation can be effectively used to improve language modeling for automatic meeting transcription. Furthermore, we have obtained some encouraging results in applying source clustering and dimensionality reduction to make more effective use of this data. In future work, we plan to exploit other sources of metadata such as e-mails, as well as the structure of the meetings themselves.

7 Acknowledgements

This research was supported by DARPA grant NG CH-D-03-0010. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

References

S. Banerjee and A. I. Rudnicky. 2007. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *Proceedings of the 2007 International Conference on Intelligent User Interfaces*, January.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of UAI'99*, Stockholm.

G. Ji and J. Bilmes. 2004. Multi-speaker language modeling. In *Proceedings of HLT-NAACL*.

S. M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401.

J. Kominek and A. Black. 2004. The CMU Arctic speech databases. In *5th ISCA Speech Synthesis Workshop*, Pittsburgh.

R. Kuhn and R. De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 570–583.

N. Morgan, D. Baron, S. Bhagat, R. Dhillon H. Carvey, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. Meetings about meetings: research at ICSI on speech in multiparty conversation. In *Proceedings of ICASSP*, Hong Kong, April.

K. Rose. 1998. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. In *Proceedings of the IEEE*, pages 2210–2239.