INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 666-2900 • FAX (510) 666-2956

Language Modeling in the ICSI-SRI Spring 2005 Meeting Speech Recognition Evaluation System

Özgür Çetin† **Andreas Stolcke**†‡
{ocetin,stolcke}@icsi.berkeley.edu

†**International Computer Science Institute, Berkeley, CA**
‡**SRI International, Menlo Park, CA**

TR-05-006

July 2005

Abstract

In this report, we describe the language models (LMs) used in the ICSI-SRI system for the NIST Spring 2005 Meeting Rich Transcription (RT-05S) evaluation. Our LMs are linear interpolations of n-gram models trained on a small number of in-domain sources and a large number of out-of-domain sources, which include conference proceedings and newly collected web data, in addition to other commonly-used corpora. Despite the lack of any training data for the lecture recognition task in the evaluation, effective LMs for this task are designed. As compared to the LMs of the ICSI-SRI-UW system for the NIST Spring 2004 Meeting Rich Transcription (RT-04S) evaluation, significant improvements in perplexity and word error rate (WER) are obtained, which are mainly due to the additional training data from the web and conference proceedings.

1 Introduction

Recent years have seen a surge of interest in recognizing speech from multi-party interactions; a sizable amount of meetings have been recorded at various sites including CMU [19], ICSI [13], and NIST [4], and a large data collection effort is currently undertaken by the Augmented Multi-party Interaction (AMI) program [1]. There is also an ongoing effort to record lectures and seminars by the Computers in the Human Interaction Loop (CHIL) consortium [2]. In May 2005, NIST has conducted an evaluation of meeting speech recognition systems, with test data drawn from both meetings (contributed by ICSI, CMU, NIST, Virginia Tech, and the AMI program) and lectures (contributed by the CHIL consortium) [3].

Speech from multi-party interactions such as meetings and lectures presents a number of unique challenges over speech from monologues and dialogs, which until quite recently have been the focus of speech and language processing research. Speaker interruptions, overlaps, reverberation, and distance microphone recordings need to be all addressed for robust speech recognition as well as for speaker tracking and segmentation [13].

One of the main challenges for language modeling in new applications in general and in the present meeting and lecture speech recognition tasks in particular is data sparsity, as accurate language models require very large training texts. Yet, limited or no training data is available for new applications typically , due to high costs associated with data collection and labeling. Model adaptation is an effective method for dealing with data sparsity, where an existing model is adapted to a new domain by re-estimating relatively few parameters. Linear interpolation of language models is a simple but powerful language model adaptation method, where predictions of several models, some from in-domain sources and some from out-of-domain sources, are linearly combined:

$$p(w|h) \equiv \sum_{s \in S} \alpha_s \, p_s(w|h). \tag{1}$$

In Equation 1, w and h denote the predicted word and its history, respectively, S is the set of sources, α_s is the interpolation weight of the s-th source, and $p_s(w|h)$ is the corresponding model. Only a few sources in Equation 1 are usually from the target domain. We use static mixture weights, i.e. the weights do not depend on word history or any other variable. The weights are the adaptation parameters, which can be estimated from a relatively small amount of in-domain data.

In addition to model adaptation, unconventional sources of data such as web [7] and written texts, e.g. conference proceedings [12], can be very helpful in sparse data situations, as we demonstrate for the meeting and lecture recognition tasks.

Our RT-05S LMs are such linear interpolations of the standard n-gram models trained on a variety of sources, including web data and conference proceedings. Due to the significant differences between the meetings and lectures in terms of speaking style, format, and content, we have trained separate LMs for these two tasks. In the following sections, we will provide detailed information about these models and our design choices. The organization of this report is as follows. We describe our training and testing data sets in the next section, followed by a description of vocabularies in Section 3. We give perplexity and WER results in Section 4 and finally present a discussion of our findings and open questions in Section 5.

2 Data

2.1 Meeting Recognition System

We have used the transcripts of 14 CMU, 69 ICSI, 11 NIST, and 28 AMI meetings for estimating in-domain LMs for meetings. A separate set of 5 CMU (24K words), 4 ICSI (35K words), 4 NIST (31K words), and 7 AMI (38K words) meetings is set aside as held-out data for estimating interpolation weights. The held-out meetings are the cluster centers in separate agglomerative clusterings of meetings from each source, using a cross-validation likelihood criterion.[1] The transcripts of the following widely-used corpora are also used: Switchboard (including Callhome, Switchboard Credit

[1] We define the distance $J(i,j)$ between the meetings i and j to be $J(i,j) \equiv \frac{1}{2}\left[D(\hat{p}_i \,||\, p_j) + D(\hat{p}_j \,||\, p_i)\right]$, where $\hat{p}_i$ is the unigram word frequency count of the meeting i, p_i is the unigram LM estimated from the transcripts of the meeting i with modified Kneser-Ney smoothing [9], and $D(\hat{p}_i \,||\, p_j) \equiv -\sum_{w \in \mathcal{W}} \hat{p}_i(w) \, \log p_j(w)$ is the cross-entropy of $\hat{p}_i$ and p_j, $\mathcal{W}$ being the LM vocabulary. Notice that $D(\hat{p}_i \,||\, p_j)$ is in fact equal to the normalized negative log-likelihood of the meeting i with respect to the model p_j. Also, $J(i,j)$ is symmetric and nonnegative, as desired. The center of a cluster of meetings is defined to be the meeting for which the total sum of distances from each meeting in the cluster to that meeting is smallest. Similarly, the cost of a cluster is defined to be the total sum of distances from each meeting in the cluster to the cluster center.

Table 1: LM training sources for the meeting and lecture tasks. Not all sources are used for both tasks (see Table 2). The corpora names under the web texts slot correspond to the corpora for which the web texts are collected.

Training source	# of words (in kilos)
Meetings	1029
AMI	144
CMU	98
ICSI	699
NIST	88
Switchboard	6558
Fisher	23357
Hub4-LM96	130850
TDT4	11869
Proceedings	28321
TED	98
Web texts	918030
AMI meetings	120149
CMU, ICSI, and NIST meetings	147510
CHIL lectures	120087
Fisher conversations	530284

Card, and Switchboard Cellular), Fisher, Hub4-LM96, and TDT4. In addition, the data collected from web that are similar to (a) CMU, ICSI, and NIST meetings [17], (b) AMI meetings, and (c) the Fisher corpus [8], are used. The AMI web data is newly collected using the recipe and tools provided in [5]. See Table 1 for a summary of our training sources.

To gain some insight into the effectiveness of various training sources on the meeting task, we calculated the perplexities for various interpolated LMs when the sources are incrementally added to the interpolation. The perplexities on the 2004 NIST meeting development set (DEV-04 consisting of 20.6K words from CMU, ICSI, LDC, and NIST meetings) and on the 2005 AMI development set (DEV-05 AMI; 62.6K words) are reported in the second and third columns of Table 2. The interpolation weights are estimated on the complete held-out data set. We find that the new AMI training data has no benefit on top of the existing meetings for DEV-04 (which does not include any AMI data), but it is very helpful for DEV-05 AMI. This hints that AMI meetings are significantly different from meetings from other sources, which is corroborated by the fact that the AMI meetings are scenario based and highly constrained topically with their own vocabularies. We also observe that the remaining sources show similar improvements for the CMU, ICSI, LDC, NIST, and AMI meetings, with most significant contributions coming from the conversational telephone speech (CTS) sources and the web data similar to the non-AMI meetings. The web sources overall reduce the perplexity on DEV-04 and DEV-05 AMI by 10% and 7%, respectively, relative.

For testing, we use the 2004 meeting evaluation set (EVAL-04) consisting of 2 CMU, 2 ICSI, 2 LDC, and 2 NIST meetings, and the aforementioned 2005 AMI development set (DEV-05 AMI) consisting of 10 AMI meetings. See Table 3 for a summary of these sets.

2.2 Lecture Recognition System

No CHIL training data was made available prior to the evaluation, other than the five lectures distributed as the development set (DEV-05 CHIL in Table 3). We have used this data for testing and adaptation in a 5-fold cross-validation fashion, where in each fold a lecture is set aside for testing while the remaining ones are used for estimating the interpolation weights. All of the training corpora used for the meetings task are also used for the lectures task, except the Fisher web data which is found to be ineffective on top of the other web data.

The CHIL lectures are generally about speech and language processing. For example, one lecture in the development set is (ironically) about language model adaptation, and another one is about speech signal processing. To

Table 2: The perplexities of various interpolated trigram models on the test sets DEV-04, and DEV-05 AMI and CHIL. The test set DEV-04 includes meetings from CMU, ICSI, LDC, and NIST. The sources are incrementally added; N/A denotes *a priori* exclusion of sources (the TED database and proceedings collection for meetings, and the TDT4 corpus and Fisher conversational web data for lectures), for which the improvements in the preliminary experiments were found to be significantly smaller than those for the included sources.

Interpolation sources	Meetings		Lectures
	DEV-04	DEV-05 AMI	DEV-05 CHIL
TED	N/A	N/A	432.1
+Meetings (Non-AMI: CMU, ICSI, NIST)	135.1	159.3	262.2
+Meetings (AMI)	135.2	115.1	257.8
+Switchboard	116.0	106.6	253.9
+Fisher	107.7	102.9	251.1
+Proceedings	N/A	N/A	160.4
+TDT4	106.3	101.5	N/A
+Hub4-LM96	100.4	95.7	157.8
+Web (Non-AMI Meetings)	90.9	90.8	150.5
+Web (AMI Meetings)	89.2	89.4	148.9
+Web (CHIL Lectures)	N/A	N/A	146.7
+Web (Fisher Conversations)	88.0	88.7	N/A

partially compensate for the lack of in-domain training data for CHIL lectures, we used the papers from major speech conferences and workshops of the last decade as training data (suggested by [12]). While such written texts are obviously very different from the oral lectures in style, they provide topic and word coverage. In addition, the transcripts of oral presentations in the Translanguage English Database (TED) are used as another source, which is expected to be closer to the CHIL lectures in style. Additional data is collected from web by using the conjunctions of frequent four-gram pairs from the Fisher corpus and the proceedings collection as the search queries, which we hoped will retrieve pages relatively similar to the lectures in terms of both style and coverage. Anecdotally, we find class discussion groups, tutorials, and technical discussion forums among the retrieved web pages.

In the last column of Table 2, we report the trigram model perplexities of various language models on DEV-05 CHIL, when sources are incrementally added to the interpolation. We find that the most significant perplexity improvements are from the addition of meeting transcripts and the proceeding papers. We find that the CTS sources do not help as much as they do for the meeting task, which is somewhat expected due to the fact that the CHIL lectures are dominated by a single speaker and that they are highly technical. The web texts also are not as helpful, but they still bring a sizable reduction (7% relative) in perplexity.

3 Vocabulary

The meetings system vocabulary consists of $54,524$ words, comprising all words in the SRI CTS system (which include all Switchboard and non-singleton Fisher words) and the ICSI, CMU, and NIST training transcripts, and all non-singleton words in the AMI training transcripts. An additional 3,483 multi-words are included in the bigram and trigram LM vocabularies but not in the 4-gram ones, to better model pronunciations of frequent n-grams. The resulting out-of-vocabulary rate on EVAL-04 is 0.32%, and 0.16% on DEV-05 AMI.

The lectures system vocabulary consists of $58,419$ words, $54,524$ of which are inherited from the meetings system vocabulary. The remaining words are the frequent words of the proceedings collection (cf. Table 1), excluding the words already in the meetings vocabulary, abbreviations, British pronunciations, and other problematic words. Similar to the meetings system, bigram and trigram LM dictionaries include an additional $3,483$ multi-words. The resulting out-of-vocabulary rate on DEV-05 CHIL is 0.18%.

Table 3: The testing sets for the meeting and lecture recognition tasks; see text for details. The subsets given in the DEV-05 CHIL slot are the individual lectures, with the original names of the data distribution.

Testing Set	# of words (in kilos)
EVAL-04	5.3
CMU	1.6
ICSI	0.7
LDC	2.0
NIST	1.0
DEV-05 AMI	62.6
DEV-05 CHIL	24.0
2004_11_11_A	3.2
2004_11_11_B	6.1
2004_11_11_C	8.3
2004_11_12_A	2.4
2004_11_12_B	4.0

4 Experiments

We have used the SRI Language Modeling Toolkit [16] for training and testing n-gram LMs, estimated with modified Kneser-Ney smoothing [9]. To reduce the computational cost during decoding, the interpolated models are pruned with entropy-based pruning[2] [15], but the interpolation is done with unpruned models. All perplexity and WER results reported below are obtained with pruned LMs.

The decoding experiments are conducted with the ICSI-SRI RT-05S evaluation system using the acoustic data from close-talking microphones [18]. There are two sets of acoustic models using perceptual linear prediction and mel-frequency cepstrum features, which are adapted from SRI's CTS system to the meeting domain. The decoding uses the three-stage decoding structure of SRI's NIST Fall 2004 CTS Rich Transcription (RT-04F CTS) evaluation system with Tandem/HATs neural-networks features [20]. It has many stages including speaker adaptation, lattice generation, consensus decoding, n-best list rescoring, and cross-adaptation. For a full description, see [20, 18]. The system uses three language models, a bigram, a trigram and a 4-gram. The bigram LM is used for lattice generation; the trigram LM is used for decoding from lattices; and the 4-gram LM is used for lattice and n-best list rescoring. The WERs reported below are the WERs at the final stage of the decoding process, and as such, they evaluate all three n-gram models together.

4.1 Meetings

Similar to the previous work in [17, 10], we found very small perplexity reductions from source-specific LMs for CMU, ICSI, and NIST meetings, and therefore, they are treated as a single source for LM training purposes. We explored the same question for the AMI meetings vs. the non-AMI (CMU, ICSI, and NIST) meeting sources, to see whether or not all meetings sources can be uniformly handled for language modeling purposes. Three sets of language models are trained. They differ from each other by the held-out data used for estimating interpolation weights: those estimated on the combined CMU, ICSI, and NIST meetings, those on the AMI meetings, and those on all of the held-out data. The held-out data in each case is folded back into the training data of the corresponding source-specific LM, after the estimation of interpolations weights. The estimated weights for each corresponding 4-gram LM are displayed in Table 4.

We find in Table 4 that the meeting LM corresponding to the source on which the mixtures are optimized receives a large weight ("Non-AMI" and "AMI" lines). However, this weight is much larger for the AMI-optimized model than the non-AMI optimized one, suggesting that AMI meetings are significantly different from meetings from other sources. This suggestion is in line with a similar observation made in Section 2.1. The latter non-AMI interpolation

[2]The fixed pruning entropy gains of 10^{-8} for bigrams, $5 \cdot 10^{-9}$ for trigrams, and $5 \cdot 10^{-10}$ for 4-grams are used.

Table 4: The interpolation weights of 4-gram models, estimated on subsets of the held-out data. "Non-AMI" refers to all meeting sources excluding AMI, which are CMU, ICSI, and NIST meetings.

Held-out meetings	Meetings		Swbd	Fsh	TDT4	LM96	Web texts		
	Non-AMI	AMI					Non-AMI	AMI	Fsh
Non-AMI	0.41	0.06	0.04	0.14	7e-5	0.03	0.19	0.05	0.09
AMI	0.09	0.62	0.02	0.08	6e-5	0.01	0.05	0.08	0.06
All	0.32	0.20	0.03	0.13	6e-5	0.02	0.16	0.05	0.08

Table 5: The 4-gram perplexities (in **bold**) and percent WERs (in *italics*) of the mixture LMs optimized on various subsets of the held-out data (cf. Table 4) and the RT-04F meeting and RT-04S CTS LMs, on the test sets EVAL-04 and DEV-05 AMI.

Language model	EVAL-04										DEV-05	
	All		CMU		ICSI		LDC		NIST		AMI	
Held-out meetings:												
Non-AMI	**91.9**	*27.9*	**110.9**	*32.5*	**58.8**	*21.4*	**93.2**	*34.9*	**90.1**	*20.1*	**93.6**	*37.3*
Non-AMI (no web)	**102.5**	*28.9*	**120.8**	*33.4*	**65.6**	*22.0*	**101.4**	*36.0*	**110.3**	*21.5*	**99.9**	*38.4*
AMI	**118.8**	*30.1*	**146.6**	*34.7*	**72.7**	*22.5*	**122.4**	*38.1*	**112.5**	*22.2*	**84.7**	*37.3*
All	**95.4**	*28.5*	**115.8**	*33.1*	**60.2**	*21.4*	**97.1**	*35.8*	**92.7**	*20.9*	**87.0**	*36.8*
RT-04S Meeting	**97.0**	*28.7*	**116.6**	*33.1*	**61.7**	*22.0*	**97.0**	*35.7*	**99.2**	*21.1*	**106.6**	*38.3*
RT-04F CTS	**95.4**	*28.7*	**110.8**	*32.1*	**78.0**	*24.5*	**85.3**	*34.1*	**108.8**	*21.5*	**113.0**	*39.8*

has a more uniform weight distribution, which is even more so in the interpolation optimized on all of the held-out data. This can be attributed to the heterogeneity of the CMU, ICSI, and NIST meetings in general, or inter-source variabilities across training and held-out meetings. Also, we find that the magnitudes of weights of training sources highly correlate with the perplexity reductions in Table 2.

The EVAL-04 and DEV-05 AMI WERs and four-gram perplexities of these LMs are reported in Table 5. For comparison, we also report the performances of the ICSI-SRI-UW Spring 2004 meeting evaluation (RT-04S) [17] and the SRI Fall 2004 CTS evaluation (RT-04F) LMs [20].[3] The RT-04F CTS and RT-04S meeting LMs are trained on the same training sources, except that (1) the RT-04S meeting LMs did not have access to the AMI training transcripts and the newly collected AMI web data, and they utilized only a subset of the complete Fisher transcripts (whatever available as of Spring 2004) and about one-third of the Fisher web data, and that (2) the RT-04F CTS LMs do not incorporate any meeting transcripts or the web data matched to the meetings. To probe the utility of the web data, we also report the performance of LMs trained on all but web sources and optimized on the non-AMI held-out data ("Non-AMI (no web)" line in Table 5).

We first compare the newly trained LMs among themselves. First, in terms of perplexity, the LMs adapted to the non-AMI meetings and to the AMI meetings perform best on the corresponding test sets (EVAL-04 and DEV-05 AMI, respectively), which is expected. The LM adapted to all meetings is a compromise between the LMs adapted to the subsets, with small degradation in perplexity. Second, in terms of WER, we find that while the LMs adapted to the non-AMI meetings perform the best on the matched test set EVAL-04 (2.2% and 0.6% absolute better than the AMI-adapted LMs, and the LMs adapted to all sources, respectively), the AMI-adapted LMs do not show any improvement over the non-AMI adapted LMs and in fact perform 0.5% absolute worse on DEV-05 AMI than the LMs adapted to all meetings. (Due to the superior performance of the non-AMI adapted LMs on the non-AMI meetings on EVAL-04, we have chosen these LMs to be used for the RT-05S meeting evaluation, for *all* meeting sources.) Third, we find that the web data is highly effective, reducing perplexity about 5–10% relative, and WER about 1% absolute.

[3] Strictly speaking, the perplexities of the RT-04F CTS and RT-04S meeting LMs are not directly comparable to the newly trained LMs, because the former ones have smaller vocabularies (47,906 and 49,881, respectively, words vs. 54,524 words). However, the differences are not large enough, so the perplexities in Table 5 are comparable. For example, the significantly smaller perplexity of the RT-04F CTS LM on the LDC data cannot be attributed to the smaller vocabulary of this LM, as the LM adapted to the all but AMI held-out data achieves a perplexity of 92.7 (as opposed to 93.2 originally) on this data, when its vocabulary is restricted to the CTS LM vocabulary.

Table 6: The interpolation weights of 4-gram models estimated by 5-fold cross-validation on DEV-05 CHIL, where we identify each fold the lecture set aside for testing.

CV fold	Meetings		TED	Swbd	Fsh	Procs	LM96	Web texts		
	Non-AMI	AMI						Non-AMI	AMI	CHIL
2004_11_11_A	0.21	0.14	0.04	0.01	0.01	0.29	2e-3	0.08	0.05	0.18
2004_11_11_B	0.18	0.14	0.05	0.01	0.01	0.28	2e-3	0.10	0.05	0.18
2004_11_11_C	0.19	0.14	0.07	0.01	0.01	0.40	3e-3	0.05	0.04	0.10
2004_11_12_A	0.19	0.14	0.04	0.01	0.01	0.30	2e-3	0.07	0.06	0.18
2004_11_12_B	0.20	0.13	0.05	0.01	0.01	0.31	2e-3	0.08	0.05	0.17

Table 7: The 4-gram perplexities (in **bold**) and percent WERs (in *italics*) on the subsets of DEV-05 CHIL obtained by 5-fold CV (cf. Table 6), with and without web texts; and those for the RT-04F CTS and RT-04S meeting LMs.

Language model	All		2004_11_11_A		2004_11_11_B		2004_11_11_C		2004_11_12_A		2004_11_12_B	
5-fold CV	**148.3**	*27.0*	**133.2**	*34.5*	**131.5**	*24.3*	**167.8**	*22.3*	**167.3**	*30.2*	**139.8**	*33.1*
5-fold CV (no web)	**155.0**	*27.6*	**135.1**	*35.3*	**131.1**	*24.8*	**187.9**	*23.4*	**169.2**	*30.8*	**142.6**	*32.8*
RT-04S Meeting	**212.2**	*31.5*	**219.1**	*42.0*	**205.9**	*31.6*	**215.4**	*24.3*	**247.2**	*35.9*	**191.7**	*35.8*
RT-04F CTS	**320.4**	*37.6*	**331.2**	*50.3*	**361.3**	*39.3*	**293.3**	*28.8*	**375.9**	*42.0*	**283.0**	*41.1*

Comparing to the RT-04S meeting LMs, we find that the additional AMI meeting and Fisher transcripts, and the new web data reduce WER about 0.6–1% absolute and perplexity about 5% relative on EVAL-04 and 12% relative on DEV-05 AMI. The improvements are largest for the AMI meetings, due to the previously mentioned differences between the AMI and non-AMI meeting sources. Comparing to the RT-04F CTS LMs, we find that the RT-05S meeting LMs perform superior on the ICSI, NIST, and AMI meetings (3.1%, 1.4%, and 2.5%, respectively, absolute WER), but interestingly the RT-04F CTS LMs fare significantly better on the CMU and LDC meetings in terms of WER, and on the LDC meetings in terms of perplexity. This could be due to the sparsity of training data for these sources (as compared to ICSI meetings, for example), or to intra- and inter-source variability in the meetings. The training sources used for the RT-04S meeting LMs are not a superset of those used for the RT-04F CTS LMs (a smaller Fisher conversational web data collection and the incomplete Fisher transcripts are used in the RT-04S meeting LMs), and as such, the improvements of the RT-04S meeting LMs on the ICSI, NIST, and AMI meetings over the RT-04F CTS LMs are smaller, and its degradations on the CMU and LDC meetings are larger (as compared to the newly trained LMs).

4.2 Lectures

We used the 2005 CHIL development set DEV-05 CHIL in a 5-fold cross-validation fashion, where in each fold one lecture is set aside for testing, while the four remaining lectures are used for estimating the interpolation weights. We have also experimented with training CHIL-source LMs in each fold, but no perplexity improvement from such an approach is achieved. The interpolation weights for each fold are displayed in Table 6, where we identify each fold by the lecture set aside for testing. The corresponding four-gram perplexities and WERs are reported in Table 7. For comparison, we again produce the performances of the RT-04F CTS and RT-04S meeting LMs.

In terms of weights assigned to different sources, we find that the proceeding papers, meeting transcripts, and the web data receive almost all of the weight, roughly equally. Among the web texts, the texts matched to the CHIL lectures receive most of the weight, as expected. The TED transcripts also receive a sizable weight. The individual lectures are similar except that in the fold corresponding to the lecture 2004_11_11_C, a significantly higher weight is assigned to the proceedings collection.[4]

[4]We found the main distinguishing characteristics of the lecture 2004_11_11_C to be that it includes many spoken equations, for which no training source is a good match. The presence of this lecture in the training divisions of other cross-validation folds and its absence in the training division of this fold could be the cause of the significantly higher interpolation weight assigned to the conference proceedings LM in this fold. This could be due to the fact that the removal of spoken equations makes the oral lectures better matched to the written papers where the equations are only displayed and not a part of the lexical text.

In terms of perplexity, we find that the cross-validation adapted LMs perform significantly better, 30% and 54% relative improvements over the RT-04S meeting and RT-04F CTS, respectively, LMs. The smaller relative improvement over the RT-04S meeting LMs shows that the meeting transcripts are highly useful for the lectures task, probably due to the fact the ICSI meetings are dominated by the speech and language processing and other technical (e.g. computers) topics despite obvious differences in speaking style and format. The WER improvements are as impressive, 4.5% and 10.6% absolute over the RT-04S meeting and RT-04F CTS, respectively, LMs. This is again mainly due to adaptation to the CHIL data. The individual lectures show wide variation in terms of both WER and perplexity. Furthermore, we find that the perplexity and WER are quite poorly correlated, e.g. compare the lectures `2004_11_12_A` and `2004_11_12_B`. The web data on average provides a significant 0.6% absolute reduction in WER, improving the recognition performance on all lectures except the lecture `2004_11_12_B`. Intriguingly, we find that the web data do not provide any significant perplexity reduction for all lectures but the lecture `2004_11_11_C`, for which the perplexity is reduced by 11% relative. Also, these improvements are smaller than those observed for the meetings task (cf. Table 5), a possible explanation of which is that some of the lecture-relevant material extracted from the web is already available in the conference proceedings.

The RT-04S evaluation LMs used for the lecture recognition task of the evaluation are identical to the cross-validation LMs of Table 6, but they are optimized on the full CHIL development set DEV-05 CHIL.

5 Discussion

In this report, we have described the language modeling for the ICSI-SRI RT-05S evaluation system. Our LMs are standard LMs adapted to the meeting and lecture domains by linear interpolation. In addition to the meeting transcripts and the widely used corpora (e.g. Hub5 and Hub4 transcripts), we have collected new web data and exploited the conference proceedings for the lecture recognition task. The additional meeting and Fisher transcripts and the new web data brought about 1% absolute WER reduction in both the existing meetings and newly added AMI meetings. The improvements are much larger for the lectures (about 4.5% absolute WER), naturally given that the style and topic coverage of lectures are significantly different from the meetings; therefore, adaptation is crucial.

Overall, our results are encouraging that it is possible, with minimal effort, to utilize existing corpora from other genres of speech and port existing models to new domains, where there is little in-domain data (such as the AMI meetings) or even no data at all (such as the CHIL lectures). Our results also demonstrate that unconventional sources such as the web and written text sources can be very useful. However, a few issues still remain. First, the meetings domain itself is varied, with many variations in speaking style, format, and coverage. Indeed, we found that a CTS LM adapted to a particular type of meetings can fare worse than the original, unadapted LM on a different kind of meetings (see CMU and LDC columns in Table 5). The questions of when to adapt, which data to adapt to, and how to assess similarities between different sources are largely open. Second, there could be potential in more sophisticated adaptation methods than the linear interpolation [6]. On-line adaptation of language models to a particular domain or speaker, similar to speaker adaptation in acoustic modeling, could also be especially helpful given that meetings tend to draw significant amounts of speech from individual speakers. Also, it is reasonable to expect speaker overlap during training and deployment of a typical meeting recognizer (as in, for example, ICSI meetings). Third, we believe that the web and other written and electronic sources have a much larger potential for language modeling, but better search, retrieval, and relevance assessment methods are necessary to fully utilize their potential. Fourth, explicit modeling of multi-party interactions and some of the syntactic (e.g. dialogs) and semantic (e.g. topics) structure in meetings and lectures could be also beneficial, for example, via speaker- or topic-dependent language modeling which we are currently exploring. However, we have so far seen little benefit from topic-dependent models using latent semantic analysis on the meetings task [14].

6 Acknowledgments

We thank Ivan Bulyko for help with LM preparation and web data collection, LIMSI for scripts for processing conference proceedings, and Barbara Peskin for helpful comments on this manuscript.

References

[1] Augmented Multi-Party Interaction (AMI), http://www.amiproject.org/.

[2] Computers in the Human Interaction Loop (CHIL), http://chil.server.de/.

[3] NIST Spring 2005 Meeting Recognition Evaluation, http://www.nist.gov/speech/tests/rt/rt2005/spring/.

[4] NIST Meeting Room Project, http://www.nist.gov/speech/test_beds/mr_proj/.

[5] Web Data for Language Modeling, http://ssli.ee.washington.edu/people/bulyko/WebData/.

[6] J.R. Bellegarda, "Statistical Language Model Adaptation: Review and Perspectives," *Speech Communication*, vol. 13, pp. 93–108, 2003.

[7] I. Bulyko, M. Ostendorf and A. Stolcke, Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-dependent Mixtures, In *Proc. of Human Language and Technology*, vol. 2, pages 7–9, 2003.

[8] I. Bulyko, M. Ostendorf and A. Stolcke, Class-dependent Interpolation for Estimating Language Models from Multiple Text Sources, Technical Report UWEETR-2003-0000, University of Washington Department of Electrical Engineering, 2003.

[9] S. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," *Computer Speech and Language*, vol. 13, pp. 359–394, 1999.

[10] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman and S. Renals, "Transcription of Conference Room Meetings: An Investigation," In *Proc. of Interspeech*, 2005.

[11] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke and C. Wooters, "The ICSI Meeting Corpus," In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pages 364–367, 2003.

[12] L. Lamel, G. Adda, E. Bilinski and J.L. Gauvain, "Transcribing Lectures and Seminars," In *Proc. of Interspeech*, 2005.

[13] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke and C. Wooters, "Meetings about Meetings: Research at ICSI on Speech in Multiparty Conversations," In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, pages 740–743, 2003.

[14] M. Pucher and Y. Huang, "Latent Semantic Analysis based Language Models for Meetings," *Proc. of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[15] A. Stolcke, "Entropy-based Pruning of Backoff Language Models," In *Proc. of DARPA Broadcast News Transcription and Understanding*, 1998.

[16] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," In *Proc. of Intl. Conf. on Spoken Language Processing*, vol. 2, pages 901–904, 2002.

[17] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf, "Progress in Meeting Recognition: The ICSI-SRI-UW Spring 2004 Evaluation System," *Proc. of NIST RT-04 Meeting Recognition Workshop*, 2004.

[18] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, 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," *Proc. of NIST RT-05 Meeting Recognition Workshop*, 2005.

[19] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu and K. Zechner, "Advances in Automatic Meeting Record Creation and Access," In *Proc. of Intl. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pages 597–600, 2001.

[20] Q. Zhu, A. Stolcke, B. Chen and N. Morgan, "Incorporating Tandem/HATs MLP Features into SRI's Conversational Speech Recognition System,", In *Proc. of DARPA RT-04F Rich Transcription Workshop*, 2004.