Adaptive Language Modeling With Varied Sources to Cover New Vocabulary Items

Sarah E. Schwarm, *Student Member, IEEE*, Ivan Bulyko, *Member, IEEE*, and Mari Ostendorf, *Senior Member, IEEE*

***Abstract*—N-gram language modeling typically requires large quantities of in-domain training data, i.e., data that matches the task in both topic and style. For conversational speech applications, particularly meeting transcription, obtaining large volumes of speech transcripts is often unrealistic; topics change frequently and collecting conversational-style training data is time-consuming and expensive. In particular, new topics introduce new vocabulary items which are not included in existing models. In this work, we use a variety of data sources (reflecting different sizes and styles), combined using mixture n-gram models. We study the impact of the different sources on vocabulary expansion and recognition accuracy, and investigate possible indicators of the usefulness of a data source. For the task of recognizing meeting speech, we obtain a 9% relative reduction in the overall word error rate and a 61% relative reduction in the word error rate for "new" words added to the vocabulary over a baseline language model trained from general conversational speech data.**

***Index Terms*—Language modeling, mixture models, speech recognition, text normalization, varied data sources.**

I. Introduction

MANY state-of-the-art speech recognizers rely on statistical language models (LMs). These models are able to automatically capture many characteristics of spontaneous speech, but most systems need a large amount of in-domain training data, on the order of millions of words. Good performance is only achieved when the training data closely matches the test data in terms of both content (topic) and style; such "in-domain" data is expensive and time-consuming to acquire for conversational speech. Written text is much more easily available than transcribed speech, but its style is often not well-suited for training language models for conversational speech recognition. In this work, we attempt to improve speech recognition performance for a conversational task by collecting text data from a variety of sources, which we combine with a general conversational speech language model.

The focus of our work is to improve recognition of "new" vocabulary items, i.e., words that were not in the baseline language model. Simply adding words to the vocabulary of the recognition system does not work; the new words need to be included in the training data in order to have a high enough language model probability to be recognized. Thus we collected topic-matched data from a variety of sources to include these new words in our training data.

The specific task we address is automatic transcription of meetings. Since our goal is to have a system that can be used for many types of meetings, we cannot assume that we will have meeting-style training data for every possible topic. Instead, we use small amounts of meeting data from a variety of meetings, topic-specific text data, and style- and topic-specific data collected from the World Wide Web to adapt a language model from a more general, conversational-speech domain into one that can be used for the meeting transcription task. We also use automatic text normalization techniques to make the text data more closely resemble spoken language. The results are analyzed in terms of overall word error rate and word error rate on the new words, to provide insights into the usefulness of different types of data sources.

The remainder of this paper is organized as follows: Section II provides background on language modeling and other work on language model adaptation. Section III presents our general approach to this problem, with details of the target task domain given in Section IV. Section V presents an analysis of style differences between corpora. Experimental results follow in Section VI. We summarize our findings and describe future directions in Section VII.

II. Background

Statistical language models typically represent the probability of a word sequence as a product of the probability of each word given its history

$$P(w) = P(w_1)\prod_{i=2}^{m} P(w_i|w_{i-1},\ldots,w_1). \quad (1)$$

Considering the full history for each word is infeasible in practice, so truncated histories are used. This results in the most commonly used statistical language model, the n-gram model, in which it is assumed that the word sequence is a Markov process. The trigram model is a very popular language model, where each word depends only on the preceding two words (a Markov process with order 2). Thus, the probability of sequence w is given by

$$P(w) = P(w_1)P(w_2|w_1)\prod_{i=3}^{m} P(w_i|w_{i-1},w_{i-2}). \quad (2)$$

Despite its simplicity, the trigram model is very successful.

Manuscript received June 10, 2003; revised October 10, 2003. This work was supported by IBM through the Faculty Award Program and by DARPA EARS under Grant MDA972-02-C-0038. A pilot version of these results was presented in [20]. The work described in this paper is based on more accurately segmented data, additional training data sources, and more detailed analysis of the impact of these sources on vocabulary expansion, and language model training. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. G. Zweig.

S. E. Schwarm is with the Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: sarahs@cs.washington.edu).

I. Bulyko and M. Ostendorf are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: bulyko@ee.washington.edu; mo@ee.washington.edu).

Digital Object Identifier 10.1109/TSA.2004.825666

Good language models require large amounts of training data that are well-matched to the target task. In this work, we use small amounts of in-domain data to adapt language models for a more general conversational speech domain. Language model adaptation can take several forms. In this work, we look at (offline) task-level adaptation, in which the models are adapted in advance using data chosen for a particular task. This is different from unsupervised cache adaptation techniques [1]–[3], where the model changes at run-time based on the utterances that have been recognized already. Since the error rates are relatively high for recognizing speech in meetings (roughly 35–40%), having a good initial model that covers new vocabulary items is important. Hence, task-level adaptation is an appropriate choice for this domain.

Previous task-level LM adaptation efforts include adding unigram probabilities from data for the target domain to an existing class bigram [4], using part-of-speech conditioning for weighting the out-of-domain data [5], and selectively weighting out-of-domain data based on word frequency counts [6], probability (or perplexity) of word or part-of-speech sequences [7], latent semantic analysis [8], and information retrieval techniques [7], [9]. Perplexity-based clustering has also been used for defining topic-specific subsets of in-domain data [10]–[12], and test set perplexity has been used to prune less useful documents from a training corpus [13]. In this work, we do not vary the modeling or data selection methods, but rather focus on obtaining different sources and analyzing their impact in a mixture modeling framework.

In real meetings and many other potential speech transcription applications, topics change frequently, making it impossible to have enough "in-domain" transcribed speech training data for any given topic. We consider training data to be "in-domain" if it matches the test data for a particular task in terms of both content and style. For example, if we want to recognize meetings from a particular research group, in-domain training data would consist of transcripts of previous meetings from that research group. For this and many other conversational tasks, acquiring sufficient in-domain training data is prohibitively expensive, and we assume that only a small amount of such data is available, i.e., for model tuning but not n-gram training. Thus we would like to be able to use out-of-domain data sources, that may be mismatched in either topic or style, to enhance language models trained on general speech data. In this example, the out-of-domain data can come from transcripts of meetings on other topics (style-matched data), written text on the same topic as the meeting (content-matched data), or text collected automatically from the World Wide Web.

Recently researchers have turned to the World Wide Web as an additional source of training data for language modeling. For "just-in-time" language modeling [14], adaptation data is obtained by submitting words from initial hypotheses of user utterances as queries to a Web search engine. Their queries, however, treated words as individual tokens and ignored function words. Such a search strategy typically generates text of a more formal written style, hence not ideally suited for recognition of conversational speech. In [15], instead of downloading the actual Web pages, the authors retrieved n-gram counts provided by the search engine. Such an approach generates valuable statistics but limits the set of n-grams to ones occurring in the baseline model. In [16] the authors achieved significant word error rate reductions by supplementing training data with text from the Web and filtering it to match the style and/or topic of the meeting recognition task. Here, we will use these Web texts as an additional training source.

Although adding in-domain training data is an effective means of improving language models [17], adding out-of-domain data is not always successful. In particular, the use of text sources in training language models for conversational speech can sometimes degrade recognition performance [7]. Hence, a side goal of this work is development of guidelines for types of data that are useful and criteria for assessing the value of a data source. It turns out that a data source that is good for n-gram training may not be good for vocabulary expansion, and vice versa. We look at previously proposed criteria (perplexity and n-gram hit rates) in the context of overall recognition performance and performance on new vocabulary items. Recognizing these new, often domain-specific words is important because even if we cannot produce perfect transcripts on a new topic, good coverage of new vocabulary items can benefit information retrieval and extraction tasks.

III. General Approach

Given that there will never be enough transcribed speech data to build task-dependent language models for most conversational speech tasks, it is important to be able to use other data sources, which naturally would include text data. In order to make better use of out-of-domain data sources, we apply text normalization to written text, expand the vocabulary with words from topic-matched sources, and use mixture techniques to combine text and conversational speech language models, as described below.

A. Data Sources for LM Training

We consider five categories of supplemental data: 1) "published" text, which consists of hand-selected papers and Web pages relating to the meeting recorder research group; 2) email, consisting of archived mailing list messages sent to the target group mailing list and to two other related mailing lists;[1] 3) speech from meetings of groups other than the group represented in the test set; 4) conversational-style text from the Web; and 5) Web-pages related to topics similar to what was discussed in the meetings. Table I lists the size of each supplemental corpus. Hand-selection of a small amount of topic-specific text data is realistic for this task; we envision a scenario in which a group wishing to use this system could easily provide papers, memos, etc., relating to the topic of an upcoming meeting. The supplemental meeting speech closely matches in style but not so in topic, because it is associated with a different group of people dealing with a different research problem.[2] The published and email text is drawn from resources associated

[1]Ideally we would just use the target group emails, but since this was early in the project there was very little text available (only about 4000 words), so we chose to augment this with messages from somewhat more general lists.

[2]Due to the nature of the corpus (primarily meetings that occurred at ICSI), there is some speaker overlap between the training data and test data, so speaker-specific dependencies may be inadvertently captured by this approach.

TABLE I
SIZE OF CORPORA FOR TRAINING SUPPLEMENTAL LANGUAGE MODELS

Corpus	Size (# words)
Published Text	32k
Email	78k
Meetings	173k
Conversational-style Web text	61M
Topic-related Web text	28M

with the target meeting group, so it is assumed to be topic-specific training data. The Web text is selected to roughly match either style or topic, with a bias toward a more informal style.

Most of the text on the Web is nonconversational, but there is a fair amount of chat-like material that is similar to conversational speech though often omitting disfluencies. This more informal style of text was our primary target when extracting data from the Web. Queries submitted to Google were composed of n-grams that occur most frequently in the switchboard training corpus, e.g., "I never thought I would," "I would think so," etc. We were searching for the exact match to one or more of these n-grams within the text of the Web pages. Web pages returned by Google for the most part consisted of conversational-style phrases like "we were friends but we don't actually have a relationship" and "well I actually I I really haven't seen her for years."

We used a slightly different search strategy when collecting topic-specific data from the Web. First we extended the baseline vocabulary with words from the meeting data and then we used n-grams with these new words in our Web queries, e.g., "wireless mikes like," "I know that recognizer." Web pages returned by Google mostly contained technical material related to topics similar to what was discussed in the meetings, e.g., "we were inspired by the weighted count scheme… ," "for our experiments we used the Bellman-Ford algorithm…," etc. The selected topic-related data is also somewhat conversational, because these texts were extracted from newsgroups, which often feature a chat-like dialogue between participants.

B. Text Normalization

The meeting data is transcribed speech and therefore may be used directly for language model training with good results. However, text corpora are unlike transcribed speech in a variety of ways. In particular, written text also includes numbers (e.g., 101, 1/2, VII, $3M), abbreviations (e.g., mph, gov't), acronyms (e.g., IBM, NIST), and other "nonstandard words" (NSWs) which are not written in their spoken form. In order to effectively use this text for language modeling, these items must be converted to their spoken forms. This process has been referred to as text conditioning or normalization and is often used in text-to-speech systems.

Text conditioning has long been used in preparing text data for language model training, and a set of text conditioning tools are available from the Linguistic Data Consortium (LDC) [18]. The LDC tools perform text normalization using a set of ad hoc rules, converting numerals to words and expanding abbreviations listed in a table. A more systematic approach to the NSW normalization problem is introduced in [19], referred to here as the NSW tools. These tools use models trained on data from several categories: news text, a recipes newsgroup, a PC hardware newsgroup, and real-estate ads. The NSW tools perform well in a variety of domains, unlike the LDC tools which were developed for business news. Thus we hypothesized that these tools would be more appropriate for conversational speech.

The NSW tools are built on a taxonomy of 23 categories, including numeric and alphabetic labels. The alphabetic labels include: ASWD, indicating that a token should be said as a word; LSEQ, meaning that a token is read as a sequence of individual letters; and EXPN, indicating that a token is an abbreviation that should be expanded to its full form. Other tokens refer to different types of numbers (e.g., dates, money, cardinal, ordinal). The text normalization process involves first splitting complex tokens using a simple set of rules, and then classifying all tokens as one of the 23 categories using a decision tree. After a token is classified, it is expanded according to type-dependent predictors. We used the NSW tools tuned on data from the PC hardware newsgroup, since this was the most similar domain to our task of recognizing technical research group meetings. We also added 52 domain-specific abbreviation expansions after examining the output of the tools when used on our topic-specific text. We compared the output of the NSW tools and the LDC tools on our published text and email corpora. Of course, not all sentences have perfect transcriptions, but a brief inspection suggests that the NSW tools have fewer errors. In our initial work, the LDC tools result in higher-perplexity language models [20], so in this work we use the NSW tools exclusively.

The retrieved Web pages required a small amount of additional filtering prior to applying the NSW tools and using the content of the pages for language modeling. First we stripped the HTML tags and ignored any paragraphs with an out-of-vocabulary (OOV) rate greater than 50%. This threshold was chosen to filter sentences that were not in English or had large numbers of errors, without eliminating short sentences that had one out-of-vocabulary word. We then piped the text through a maximum entropy sentence boundary detector [21] and performed text normalization using the NSW tools.

C. Vocabulary Expansion

One of the main goals of this work is to improve recognition of "new" words, that is, words which are not in the baseline language model and vocabulary. In addition to collecting topic-specific training data which includes new words, we must choose a list of words to add to the vocabulary of the speech recognizer. We did this by choosing words which occur at least 5 times in one of the supplemental data sources. The 5-occurrence threshold was a simple heuristic used to avoid adding new words that were simply typos. Pronunciations for the new words were obtained from a larger dictionary when available, and generated by hand for the relatively small number of words not covered in that dictionary.

We only selected words from the supplemental sources that were closely matched to the target domain: meetings, text, and email. We chose not to add new words from the Web corpora due to the high rate of incorrect spellings as well as offensive words. For example, words "amature," "becuase,"

"definately," and "dumbass" were among the top candidates from the topic-related Web text. There were also a fair number of British spellings, e.g., "centre," "colour," etc.

Another reason for not using words selected from the Web data is that we had to add many words from the Web sources in order to get small improvements in the OOV rate. By adding 250 words from the Web sources, we only covered an additional 10 tokens in the test set. In order to get 50 hits, 1000 new words were needed. In contrast, adding 86 words from the topic-related text sources gave 131 hits. Since most of the new words from the Web sources are not actually in the test data, it is not worth the effort of adding so many of them. The closely matched sources provided much higher gains with many fewer new words.

D. Mixture Language Models

A common technique for combining several language models is a mixture model, a linear interpolation of two or more component models considered at the n-gram level [22], [23]. Mixture components can include models from different corpora, as used in this paper, or topic-dependent models trained on subsets of a particular corpus. In the trigram case, each probability $P(w_i|w_{i-1}, w_{i-2})$ in (2) is replaced with a weighted sum of probabilities from $|S|$ individual models

$$P(w_i|w_{i-1}, w_{i-2}) = \sum_{s \in S} \lambda_s P_s(w_i|w_{i-1}, w_{i-2}). \quad (3)$$

The interpolation weights λ_s are estimated automatically using the Expectation-Maximization algorithm to maximize likelihood on a small held-out data set (or, equivalently, minimize perplexity) with the constraint that $\sum_{s \in S} \lambda_s = 1$. Note that the mixture models require some in-domain training data in order to estimate the mixture weights.

In this work, we combined a baseline language model for conversational speech with supplemental LMs trained on several different text and conversational speech data sources. The baseline LM, which is also a mixture model, is described in more detail in Section IV. All language models were estimated using the SRI Language Modeling Toolkit [24] with the modified Kneser-Ney discounting scheme [25].

Combining several n-grams can produce a model with a very large number of parameters, which is costly in decoding. In such cases n-grams are typically pruned. In most of the work reported here, the models are unpruned. However, in some of the experiments involving Web data sources, the final mixtures were aggressively pruned to about 20% of their original size. We use entropy-based pruning [26] after combining unpruned models, in all cases using the same threshold (entropy gain of 10^{-8}).

IV. Task Domain and Experiment Paradigm

Our work is part of the ICSI/UW Meeting Recorder project [27], the goal of which is to develop a system for automatically transcribing and browsing meeting speech. This target task uses data collected by ICSI. Meetings in the corpus are regularly scheduled group meetings at ICSI, i.e., real meetings that would occur even if they were not being recorded for this project. This work is based on a pilot release of meeting data which comprises our test data (five meetings from the meeting recorder group), held-out data (four other meetings from this group) and style-specific data from meetings on other topics used as supplemental training data.

A. Test Sets

Our test data consists of meetings of the Meeting Recorder project group at ICSI. For the results reported here, the evaluation test set consists of five 1-hour meetings (approximately 44 000 words) from one group. We exclude speakers who are not native speakers of American English, as in [27]. We also used approximately 39 000 words of data from other meetings of this group as a held-out set for LM mixture weight estimation and optionally for pruning, and we had a separate development test set of about 56 500 words.

B. Recognizer

For our recognition experiments, we used a modified version of SRI's large-vocabulary conversational speech recognition system from the March 2000 Hub-5 evaluation [23].[3] The current system uses new acoustic models trained using MMIE, and the baseline language model, described below, has been updated since the evaluation. There were also minor modifications for the meeting task, including downsampling the meeting speech in order to use the telephone-band acoustic models from the Hub-5 system [27]. This system processes the test data in two passes. The first pass uses a relatively simple language model to generate n-best lists: lists of the n most likely hypotheses for each utterance, consisting of an acoustic score and a language model probability for each hypothesis. These lists are rescored using a more complex model. In experiments described in this paper, the first-pass recognizer used a bigram LM to generate n-best lists with $n = 1000$, followed by a rescoring pass using a trigram LM. The oracle error rate for the n-best lists was 22.7%.

C. Baseline LMs

Our baseline bigram and trigram language models were an updated version of the LMs for the SRI Hub-5 recognizer from the March 2000 evaluation, with the main changes being inclusion of new training data and consistent smoothing using the Kneser-Ney backoff. Both the bigram for the first-pass search and the trigram used in rescoring models were mixtures built from individual n-gram models trained on data from the Switchboard, CallHome, Switchboard-cellular and Broadcast News corpora. The combined Switchboard and Callhome corpora consisted of about 3 million words, and Broadcast News was 150 million words. The baseline models as well as our supplemental models use multi-words, lexical entries that contain multiple words, e.g., "you_know" and "a_couple_of." Without multi-words, the baseline vocabulary is 34 898 words, and including them it is 36 552. Both baseline mixtures were pruned using a relative entropy gain threshold of 10^{-8}.

[3]The March 2000 Hub-5 evaluation is one of a series of NIST-sponsored benchmark tests of speech recognition for conversational speech over the telephone.

TABLE II
FREQUENCY OF SELECTED WORD TYPES IN SWITCHBOARD, MEETING DATA, PUBLISHED TEXT SOURCES, EMAIL, AND WEB TEXTS DEMONSTRATING DIFFERENCES BETWEEN THESE DOMAINS

POS	Unigram Frequency (%)					
	Swbd	Mtgs	Email	Pub	Conv-web	Topic-web
Pronouns	14	10	3	2	13	10
Nouns	13	17	29	31	19	22
Adjectives	4	6	11	10	9	9
Adverbs	8	9	4	4	6	5
Filled Pauses	3	3	10^{-5}	10^{-4}	0.04	0.02
Back-channels	2	1	0	0	0.3	0.2

TABLE III
FREQUENCY OF OCCURRENCE (%) OF SPECIFIC COORDINATING CONJUNCTIONS AT THE BEGINNING OR END OF A SENTENCE

Sentence Position	CC	Frequency (%)			
		Swbd	Mtgs	Email	Pub.
start	and	20	34	2	<1
	but	24	37	10	3
	so	23	45	22	7
end	and	4	9	0	0
	but	7	16	<1	0
	so	13	18	<1	0

V. ANALYSIS OF DIFFERENCES BETWEEN CORPORA

A. Style Differences Between Corpora

The style of conversational speech differs greatly from written text. This difference can be characterized in part by variations in part-of-speech usage patterns, as illustrated in [7] with a comparison of Switchboard, Broadcast News, and Wall Street Journal data. Table II provides an analysis of selected word categories in our data, to illustrate differences in the corpora. Filled pauses are words that are typically used by the speaker to hold the floor while thinking of the next word to say, e.g., "um" and "uh." Back-channels are words like "yeah," "uh-huh," and "right" that are uttered by the listener while someone else is speaking. Both filled pauses and back-channels are relatively frequent in conversational speech, but rare in text data.[4] The pattern of more pronouns in speech and more nouns in written text is consistent with that observed in [7]. We also note that there are more coordinating conjunctions ("and," "so," "but," "or," "nor," "yet") in speech than in text. Further analysis of the location of specific coordinating conjunctions shows that certain of these words (e.g., "and," "but," "so") occur frequently at the beginning or end of utterances[5] in conversational speech, while they almost never occur at the beginning or end of sentences in written text. Table III shows the percentage of all occurrences of the most common coordinating conjunctions at the beginning or end of a sentence (in text) or utterance (in speech). Data is not included for Web sources, since sentence boundaries were tagged automatically so the numbers may not be reliable. We also analyzed the location of other coordinating conjunctions as well as filled pauses, but did not find clear patterns of occurrence for these words.

Like Switchboard, meetings often include casual, conversational speech. In many cases, participants are friends as well as colleagues. Based on the patterns seen here, we can classify Switchboard and the meeting corpus as more stylistically similar, while published text and email are more closely matched in topic but not style. The two Web corpora tend to have POS patterns that are somewhere between these extremes. Of course, meetings have different styles—e.g., formal committee meetings differ from research group brainstorming sessions—and not all styles are represented in our data. In addition, the patterns of usage of some of these conversational speech fillers can be speaker-dependent [28].

[4]Although they are typically markers of conversational speech, not text, filled pauses and back-channels have nonzero probability in the text data because the group studies conversational speech, so sometimes words like "uh-huh" and "uh" are discussed.

[5]We use the term "utterance" to denote a sentence-like segment of speech, since conversational speech often cannot be accurately divided into grammatical sentences.

TABLE IV
OUT-OF-VOCABULARY (OOV) RATES ON MEETING TEST DATA USING BASELINE VOCABULARY ALONE (36552 WORDS) AND SUPPLEMENTED WITH WORDS FROM OTHER SOURCES

LM Source	# Additional Words	OOV Rate (%)	
		Tokens	Types
Baseline	–	1.35	7.32
+ Published Text	86	1.06	6.69
+ Email	228	0.96	6.11
+ Meetings	111	1.15	6.72
+ All	413	0.80	5.50

B. Content Differences Between Corpora

Prior to building new language models for recognition experiments, we looked at the effect of adding words from closely matched supplemental sources to the baseline vocabulary (originally 36 552 words). For each supplemental data source, we selected words that occurred at least 5 times in that source (to avoid typos) but were not in the baseline vocabulary. The results tabulated in Table IV show that in all cases, the rate of occurrence of out-of-vocabulary (OOV) words was reduced. New words from the meeting corpus reduced the OOV rate by the same amount as words from published text, and almost as much as words from email (the topic-specific sources). However, the meeting corpus is ten times the size of the text corpus. By using topic-specific text, we can reduce the OOV rate with a much smaller amount of training data. Not surprisingly, adding words from all three sources yielded the greatest reduction. In this case, we added words that occurred at least 5 times across all the corpora, including 56 words that occurred in multiple corpora but occurred fewer than 5 times in any individual corpus. As discussed earlier, the Web data was not a good source of new words and therefore is not included here.

In addition to OOV rate, two other measures of source mismatch/content similarity between corpora are language model perplexity and n-gram hit rate. Perplexity is an information-theoretic measure that, put simply, characterizes the branching factor of a language model. It is often used as a quick way to assess the quality of a model, although Iyer has shown that perplexity is not always an accurate measure when out-of-domain data is used [29], [30]. N-gram hit rate is a measure of how many n-grams in the target data are actually represented in the language model. It has been suggested that n-gram hit rate might be another good way to easily assess language model quality.

TABLE V
MEASURES OF SOURCE MISMATCH ON MEETING DEVELOPMENT SET: PERPLEXITY (PP), n-GRAM HIT RATE. MODELS ARE INDIVIDUAL SUPPLEMENTAL MODELS, NOT MIXTURES (EXCEPT FOR THE BASELINE)

LM Source	PP	2gr hit	3gr hit
Baseline	120	88.3	27.8
Published Text	814	20.9	0.4
Email	564	32.8	0.7
Meetings	120	67.5	10.6
Conv Web	292	82.6	31.4
Topic Web	289	81.8	24.9

Table V shows perplexity, trigram hit rate, and bigram hit rate for the individual component language models, measured on the development set. As expected, there is a direct relationship between bigram and trigram hit rate. The hit rates also reflect the size of the data set used to train the language model. The baseline and Web LMs have the highest hit rates while the small published text and email models have the lowest. The published text and email LMs also have the highest perplexity, probably because there was so little text available for those models. There does not appear to be a high correlation between hit rate and perplexity, in that the text and email LMs have lower hit rates and higher perplexity than the baseline, while the conversational Web LM has both a higher trigram hit rate and higher perplexity and the meeting LM has a much lower hit rate but the same perplexity.

VI. EXPERIMENTAL RESULTS

For our work, we modified the baseline models by adding 413 new vocabulary entries taken from the closely-matched supplemental sources and renormalizing the model. This made no difference to baseline recognition performance, but it did affect language model perplexity. Initially, we used a bigram version of the baseline model in the first pass recognition to generate the n-best lists. Rescoring these n-best lists with a trigram LM gave an average word error rate (WER) of 39.1%. However, the WER on the new vocabulary items was very high (85.0%), more than double the overall WER. Since the new words did not occur in the training data used to generate the baseline model, these words did not have meaningful unigram probabilities assigned to them and hence were largely excluded from the n-best lists. In order to have a better "starting point" we used a bigram mixture of the baseline, text, email and meetings data sources to recompute the n-best lists, which were then rescored to produce the results reported in this section. This choice of the first pass recognition LM provided us with a better framework to assess the influence of different data sources on WERs among the new vocabulary items, although the results for mixtures where data sources did not include all of the above (i.e., baseline, text, email, and meetings) may be overly optimistic.

Recognition results for the baseline LM and all the mixture models are presented in Tables VI and VII for the full evaluation test set and the subset of tokens that correspond to new words in the vocabulary. In addition, perplexity and trigram and bigram hit rates on the respective sets are reported. Each of the individual supplemental sources provides at least a small improvement, with larger gains for the mixtures that combine multiple supplemental sources. An improvement of 3.4% absolute or 9% relative comes from using all the supplemental sources together, compared to the baseline with new words added to the vocabulary without being included in the training data (from 39.1% to 35.7% WER). There is a much larger improvement in word error rate on the new vocabulary items – a 61% relative gain between this baseline model and the mixture containing all supplemental data sources (from 85.0% to 33.3% WER).

TABLE VI
OVERALL WER RESULTS FOR RECOGNITION EXPERIMENTS, PLUS PERPLEXITY (PP) AND n-GRAM HIT RATES ON EVALUATION TEST SET. ALL MODELS EXCEPT THE BASELINE ARE MIXTURES WITH THE BASELINE AS ONE COMPONENT

LM Sources	PP	2gr hit	3gr hit	WER
Baseline	120.2	85.6	26.2	38.2
+ Text	106.9	86.3	26.3	37.7
+ Email	107.0	86.4	26.3	37.5
+ Meeting	100.8	86.6	26.7	37.2
+ Conv Web	103.5	90.8	39.9	36.8
+ Topic Web	100.8	91.0	35.2	36.7
+ Text+Email+Meeting	97.4	87.5	26.8	37.0
+ Text+Email+Meeting+Conv	90.3	91.4	40.2	35.8
+ Text+Email+Meeting+Topic	91.5	91.4	35.4	35.9
+ All	89.5	92.3	41.4	35.7

TABLE VII
EVALUATION TEST SET WER RESULTS FOR THE SUBSET OF *NEW WORD TOKENS*, PLUS PERPLEXITY (PP) AND n-GRAM HIT RATES ON THIS SUBSET. ALL MODELS EXCEPT THE BASELINE ARE MIXTURES WITH THE BASELINE AS ONE COMPONENT

LM Sources	PP	2gr hit	3gr hit	WER
Baseline	45K	0	0	57.8
+ Text	204	20.0	0	42.2
+ Email	196	17.3	0.4	43.1
+ Meeting	172	25.0	1.9	39.9
+ Conv Web	759	28.5	1.2	45.4
+ Topic Web	203	56.9	5.8	38.6
+ Text+Email+Meeting	95	38.4	2.3	36.6
+ Text+Email+Meeting+Conv	105	47.3	3.5	36.6
+ Text+Email+Meeting+Topic	81	66.5	7.7	33.3
+ All	87	67.3	7.7	33.3

The conversational Web corpus is among the most useful single sources for improving the overall WER, but it is the least useful for improving recognition of new words. In contrast, the topic Web data provides much greater gains in WER for new words, with similar improvement in overall WER. While the improvement is smaller with models from smaller corpora, the text and email data also provide significant improvement in WER on the new vocabulary items, showing that topic-matched data is most important for the recognition of new words and can be effective even in very small quantities. Recall also that the Web data was not very useful for adding new vocabulary items.

Fig. 1 shows the weights chosen for each mixture component for different n-gram orders, illustrating the relative importance of each data source. For unigrams, the style of the data matters most, as evidenced by the dominating weight of the meeting data. For bigrams and trigrams, style still counts, but the size of the data set becomes more important and the larger baseline and Web corpora are given more weight.

Tables VI and VII also show the perplexity and n-gram hit rate statistics for the mixture models. Perplexity and n-gram hit rate have the benefit of being simple and quick to calculate, so

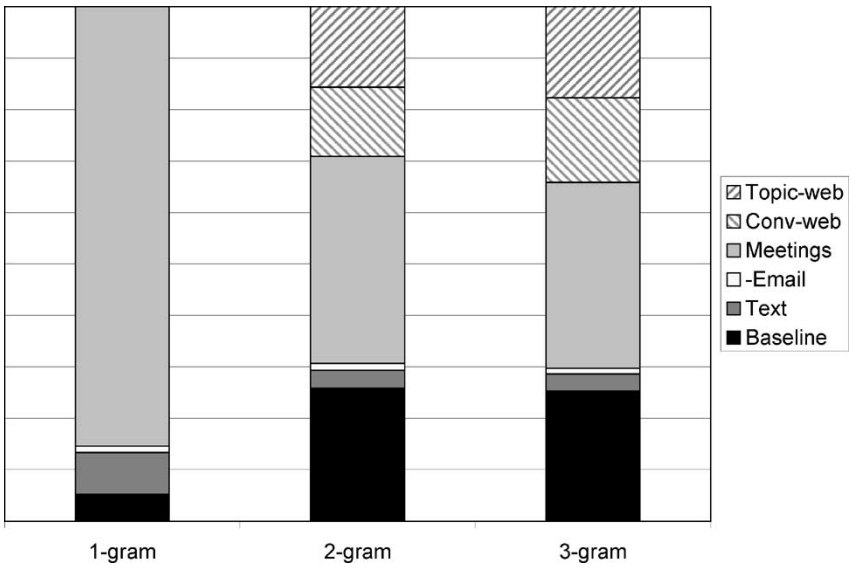

Fig. 1. Relative weight of various data sources in mixtures of different n-gram orders.

TABLE VIII
CORRELATING MIXTURE MODEL CHARACTERISTICS TO WORD ERROR RATES ON THE FULL EVALUATION TEST SET AND ON THE SUBSET OF NEW WORD (NEW-WD) TOKENS. CORRELATIONS ARE BASED ON 10 ENTRIES FOR EACH ROW

Model's characteristics on test set	Correlation with	
	WER	WER (new-wd)
perplexity	0.94	0.94
2-gram hit rate	-0.93	-0.63
3-gram hit rate	-0.86	-0.45
new-wd perplexity	0.56	0.84
new-wd 2-gram hit rate	-0.92	-0.87
new-wd 3-gram hit rate	-0.85	-0.73

we would like them to be strongly correlated with WER in order to use them as predictors of which models will perform well, without the expense of conducting recognition experiments for all possible models. We calculated correlation coefficients between these model characteristics and word error rate to further analyze the usefulness of these predictors. In Table VIII we give the correlation of overall WER (and WER on the subset of new word tokens) with three characteristics of the mixture models calculated on the test set: perplexity, bigram hit rate, and trigram hit rate. Perplexity has the strongest correlation for both overall WER and error rate on new words. The bigram hit rate is also good for overall WER, but somewhat less useful for new words. Trigram hit rate is least useful. Not surprisingly, bigram and trigram hit rates on the subset of new words are better correlated with WER on that subset, but neither is as effective as overall perplexity. Looking at the details in Tables VI and VII, it appears that trigram hit rate mainly reflects corpus size for the full vocabulary, but for the added words there is a clear impact of topic match in both bigram and trigram hit rate. Topic match also seems to matter more than size for perplexity computed only on the subset of new words, but new word perplexity is still not as useful as overall perplexity for predicting performance on the new words.

We also analyzed the dependence of WER on evaluation data with characteristics of the individual component models (from Table V, calculated on the development set). Since there are only 6 data points for each case, we illustrate these in Fig. 2 to show the relationships rather than give the correlation statistics. While there is too little data to draw strong conclusions, it appears that component-level measures are much less reliable as an indicator of potential WER reduction than measures on the complete mixture. This is not entirely surprising – it is difficult to assess impact on overall WER when looking at a component model in isolation. The new-word bigram hit rate seems to be somewhat useful for predicting performance on new words, but perplexity is not useful, even when computed only on the subset of new words. The finding that perplexity of the component model is not a good predictor may be related to the finding in [16] that perplexity-based filtering of training data does not lead to improved performance of the final system. This is not inconsistent with the prior finding that the perplexity of the *combined* model is a useful predictor, since the component model may not itself have good perplexity but could lead to improvements in combination with other models if it offers coverage of a phenomenon not well represented by the other models.

Since data collected from the Web can be huge in size, it can lead to very large language models, which are often pruned to reduce memory requirements. Hence, we also conducted experiments using pruning for all the cases involving Web data, where the size was reduced to about 20% of the original using entropy-based pruning (as described earlier). (Other data sources were so small that pruning was not necessary.) Pruning led to a small loss in performance in most cases. Using all the data, the overall WER increased from 35.7% to 36.1%, and the error rate on new words increased from 33.3% to 34.6%. Using pruning did not have a large impact on perplexity as used for model assessment, but it did make trigram hit rate effectively useless.

VII. CONCLUSION

In summary, we achieved significant reductions in overall word error rate (9% relative) and, particularly, in recognition of new vocabulary items (61% relative) by using data collected

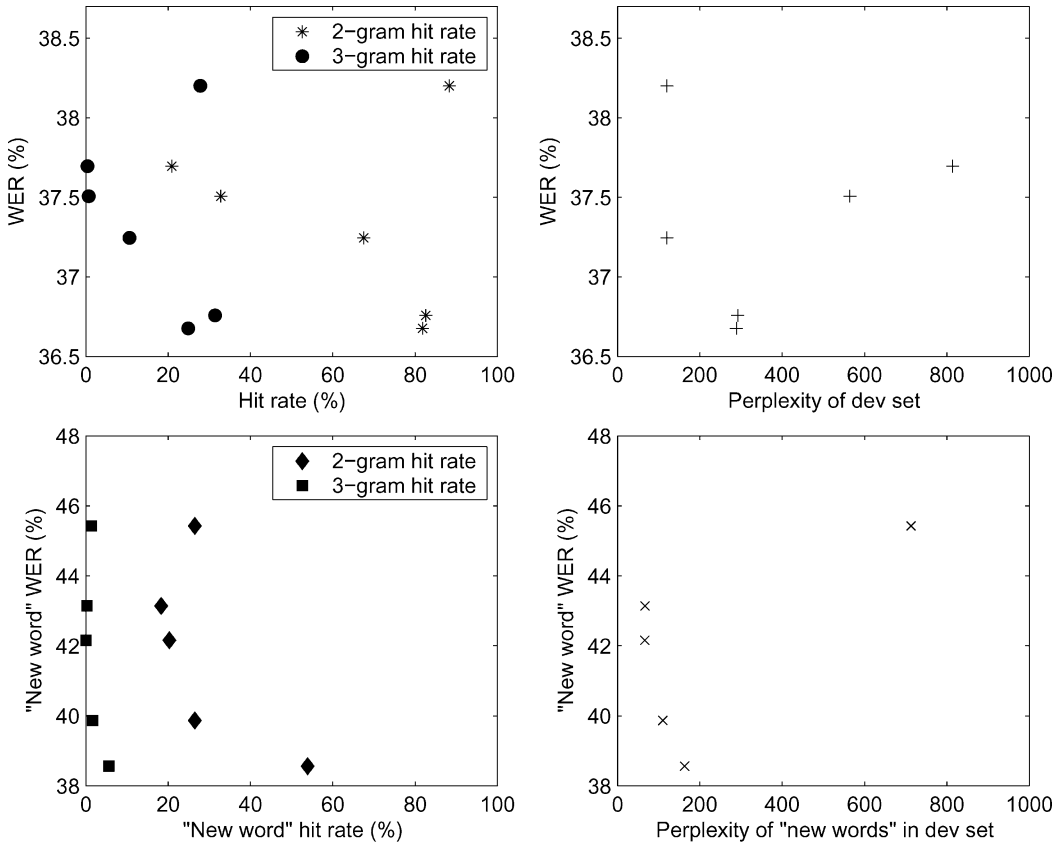

Fig. 2. Relationships between various model characteristics and WERs for the full test set (top row) and the subset of new word tokens (bottom row). The baseline model is included only in the top row, since the training for this model does not cover the new words.

from out-of-domain sources: papers, email, other meetings, and the World Wide Web. Text normalization and mixture language models were used to successfully combine these data with the baseline LM for a more general conversational speech task. Using order-dependent mixture weights, we find that the Web data is mainly useful for higher-order n-grams (i.e., not unigrams), and it is not very effective for vocabulary expansion. Larger data sources give more gain in overall performance, but topic match was more important than size for reducing WER on new words.

We also showed that perplexity can be used to assess the combined language model (but not component models) and that bigram hit rate is somewhat useful for assessing new data sources in terms of their impact on WER of targeted (new) vocabulary items.

Opportunities for future work in this area include collecting more training data from the Web and refining the existing text normalization tools. Another potential direction is to combine LMs from different domains using class-dependent interpolation [16], where a larger number of mixture weights is estimated (more than one per data source) in order to handle source mismatch, specifically letting the mixture weights vary as a function of the previous word class.

ACKNOWLEDGMENT

The authors would like to thank A. Stolcke and colleagues at ICSI for their help with recognition experiments.

REFERENCES

[1] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss, "A dynamic LM for speech recognition," in *Proc. ARPA Workshop on Speech and Natural Language*, 1991, pp. 293–295.
[2] R. Kuhn and R. de Mori, "A cache based natural language model for speech recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 570–583, 1992.
[3] A. Kalai, S. Chen, A. Blum, and R. Rosenfeld, "On-line algorithms for combining language models," in *Proc. ICASSP*, 1999.
[4] P. Witschel and H. Hoge, "Experiments in adaptation of language models for commercial applications," in *Proc. Eurospeech*, vol. 4, 1997, pp. 1967–1970.
[5] R. Iyer and M. Ostendorf, "Transforming out-of-domain estimates to improve in-domain language models," in *Proc. Eurospeech*, vol. 4, 1997, pp. 1975–1978.
[6] A. Rudnicky, "Language modeling with limited domain data," in *Proc. ARPA Spoken Language Technology Workshop*, 1995, pp. 66–69.
[7] R. Iyer and M. Ostendorf, "Relevance weighting for combining multi-domain data for n-gram language modeling," *Comput. Speech Lang.*, vol. 13, no. 3, pp. 267–282, 1999.
[8] J. Bellegarda, "Exploiting both local and global constraints for multispan statistical language modeling," in *Proc. ICASSP*, 1998, pp. II:677–680.
[9] M. Mahajan, D. Beeferman, and D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," in *Proc. ICASSP*, 1999, pp. I:541–544.
[10] R. Iyer and M. Ostendorf, "Modeling long range dependencies in languages," in *Proc. ICSLP*, 1996, pp. 236–239.
[11] P. Clarkson and A. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *Proc. ICASSP*, 1997, pp. II:799–802.
[12] S. Martin *et al.*, "Adaptive topic-dependent language modeling using word-based varigrams," in *Proc. Eurospeech*, 1997, pp. 3:1447–1450.
[13] D. Klakow, "Selecting articles from the language model training corpus," in *Proc. ICASSP*, 2000, pp. III:1695–1698.
[14] A. Berger and R. Miller, "Just-in-time language modeling," in *Proc. ICASSP*, 1998, pp. II:705–708.

[15] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," in *Proc. ICASSP*, 2001, pp. I:533–536.
[16] 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. HLT-NAACL*, Comp. Vol., 2003, pp. 7–9.
[17] R. Rosenfeld, "Optimizing lexical and n-gram coverage via judicious use of linguistic data," in *Proc. Eurospeech*, vol. 3, 1995, pp. 1763–1766.
[18] (1998) 1996 CSR Hub-4 Language Model. Linguistic Data Consortium. [Online]. Available: http://morph.ldc.upenn.edu/Catalog/LDC98T31.html
[19] R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of nonstandard words," *Comput. Speech Lang.*, vol. 15, no. 3, pp. 287–333, July 2001.
[20] S. Schwarm and M. Ostendorf, "Text normalization with varied data sources for conversational speech language modeling," in *Proc. ICASSP*, vol. I, 2002, pp. 789–792.
[21] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proc. Empirical Methods in Natural Language Processing Conference*, 1996, pp. 133–141.
[22] L. Bahl *et al.*, "The IBM large vocabulary continuous speech recognition system for the ARPA NAB news task," in *Proc. ARPA Workshop on Spoken Language Technology*, 1995, pp. 121–126.
[23] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. R. Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sommez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, May 2000.
[24] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, 2002, pp. 901–904.
[25] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, no. 4, pp. 359–394, 1999.
[26] A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 270–274.
[27] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. Int. Conf. on Human Language Technology*, 2001, pp. 246–252.
[28] E. E. Shriberg, "To 'errrr' is human: Ecology and acoustics of speech disfluencies," *J. Int. Phonetic Assoc.*, vol. 31, no. 1, pp. 153–169, 2001.
[29] R. Iyer, M. Ostendorf, and M. Meteer, "Analyzing and predicting language model improvements," in *Proc. IEEE Workshop on Speech Recognition and Understanding*, 1997, pp. 254–261.
[30] R. Iyer, "Improving and Predicting Performance of Statistical Language Models in Sparse Domains," Ph.D. dissertation, Boston Univ., Boston, MA, 1998.

Sarah E. Schwarm (S'02) received the B.A. degree in cognitive science from the University of Virginia, Charlottesville, in 1999 and the M.S. degree in computer science and engineering from the University of Washington, Seattle, in 2001. She is currently pursuing the Ph.D. degree in computer science and engineering at the University of Washington.

Her research interests are in speech recognition, natural language processing, and education.

Ms. Schwarm is a member of the Association for Computer Machinery.

Ivan Bulyko (M'99) received the B.A. degree in electrical engineering and computer science from Suffolk University, Boston, MA, in 1997, the M.S. degree in computer engineering from Boston University in 1999, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2002.

He is currently a Research Associate at the University of Washington. His research interests include speech synthesis, speech recognition, and natural language processing. His most recent work focused on improving N-gram language models of conversational English and Mandarin by obtaining additional training text from the Internet and by using class-dependent interpolation of N-grams.

Dr. Bulyko is a member of Delta Alpha Pi.

Mari Ostendorf (M'85–SM'97) received the B.S., M.S., and Ph.D. degrees in 1980, 1981, and 1985, respectively, all in electrical engineering from Stanford University, Stanford, CA.

In 1985, she joined the Speech Signal Processing Group at BBN Laboratories, where she worked on low-rate coding and acoustic modeling for continuous speech recognition. She joined the faculty of the Department of Electrical and Computer Engineering at Boston University, Boston, MA, in 1987, and since 1999 she has been a Professor of electrical engineering at the University of Washington, Seattle. Her research interests are primarily in the area of statistical pattern recognition for nonstationary processes, particularly in speech processing applications, and her work has resulted in more than 130 publications. Her early work was in speech coding; more recently she has been involved in projects on both continuous speech recognition and synthesis, as well as other types of signals. She has made contributions in segment-based and higher order acoustic models, data selection and transformation for language modeling, and stochastic models of prosody for both recognition and synthesis.

Dr. Ostendorf has served on the Speech Processing and DSP Education Committees of the IEEE Signal Processing Society and is a member of Sigma Xi.