Multi-Speaker Language Modeling

Gang Ji and Jeff Bilmes *
SSLI Lab, Department of Electrical Engineering
University of Washington
Seattle, WA 98195-2500
{gang,bilmes}@ee.washington.edu

Abstract

In conventional language modeling, the words from only one speaker at a time are represented, even for conversational tasks such as meetings and telephone calls. In a conversational or meeting setting, however, speakers can have significant influence on each other. To recover such un-modeled inter-speaker information, we introduce an approach for conversational language modeling that considers words from other speakers when predicting words from the current one. By augmenting a normal trigram context, our new *multi-speaker language model* (MSLM) improves on both Switchboard and ICSI Meeting Recorder corpora. Using an MSLM and a conditional mutual information based word clustering algorithm, we achieve a 8.9% perplexity reduction on Switchboard and a 12.2% reduction on the ICSI Meeting Recorder data.

1 Introduction

Statistical language models (LMs) are used in many applications such as speech recognition, handwriting recognition, spelling correction, machine translation, and information retrieval. The goal is to produce a probability model over a word sequence $P(W) = P(w_1, \cdots, w_T)$. Conventional language models are often based on a factorized form $P(W) \approx \prod_t P(w_t|\Phi(h_t))$, where h_t is the history for w_t and Φ is a history mapping.

The case of n-gram language modeling, where $\Phi(h_t) = w_{t-n+1} \cdots w_{t-1}$, is widely used. Typically, $n = 3$, which yields a *trigram* model. A refinement of this model is the class-based n-gram where the words are partitioned into equivalence classes (Brown et al., 1992). In general, smoothing techniques are applied to lessen the curse of dimensionality. Among all methods, modified Kneser-Ney smoothing (Chen and Goodman, 1998) is widely used because of its good performance.

Modeling conversational language is a particularly difficult task. Even though conventional techniques work well on read or prepared speech, situations such as telephone conversations or multi-person meetings pose great research challenges due to disfluencies, and odd syntactic/discourse patterns. Other difficulties include false starts, interruptions, and poor or unrepresented grammar.

Most state-of-the-art language models consider word streams individually and treat different phrases independently. In this work, we introduce *multi-speaker language modeling* (MSLM), which models the effects on a speaker of words spoken by other speakers participating in the same conversation or meeting. Our new model achieves initial perplexity reductions of 6.2% on Switchboard-I, 5.8% on Switchboard Eval-2003, and 10.3% on ICSI Meeting data. In addition, we developed a word clustering procedure (based on a standard approach (Brown et al., 1992)) that optimizes *conditional* word clusters. Our class-based MSLMs using our new algorithm yield improvements of 7.1% on Switchboard-I, 8.9% on Switchboard Eval-2003, and 12.2% on meetings.

A brief outline follows: Section 2 introduces multi-speaker language modeling. Section 3 provides initial evaluations on Switchboard and the ICSI Meeting data. Section 4 presents evaluations using our class-based multi-speaker language models, and Section 5 concludes.

2 Multi-speaker Language Modeling

In a conversational setting, such as during a meeting or telephone call, the words spoken by one speaker are affected not only by his or her own previous words but also by other speakers. Such inter-speaker dependency, however, is typically ignored in standard n-gram language models. In this work, information (i.e., word to-

*This work was funded by NSF under grant IIS-0121396.

kens) from other speakers (A) is used to better predict word tokens of the current speaker (W). When predicting w_t, instead of using $P(w_t|w_0, \cdots, w_{t-1})$, the form $P(w_t|w_0, \cdots, w_{t-1}; a_0, \cdots, a_t)$ is used. Here a_t represents a word spoken by some other speaker with appropriate starting time (Section 3). A straight-forward implementation is to extend the normal trigram model as:

$$P(w_t|\Phi(h_t)) = P(w_t|w_{t-1}, w_{t-2}, a_t). \quad (1)$$

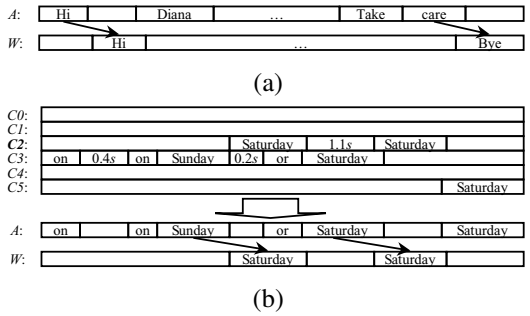

Figure 1: Examples of phone conversation (a) and meeting (b). (Frame sizes are not proportional to time scale.)

Figure 1 shows an example from Switchboard (a) and one from a meeting (b). In (a), only two speakers are involved and the words from the current speaker, W, are affected by the other speaker, A. At the beginning of a conversation, the response to "Hi" is likely to be "Hi" or "Hello." At the end of the phone call, the response to "Take care" might be "Bye", or "You too", *etc.* In (b), we show a typical meeting conversation. Speaker C2 is interrupting C3 when C3 says "Sunday". Because "Sunday" is a day of the week, there is a high probability that C2's response is also a day of the week. In our model, we only consider two streams at a time, W and A. Therefore, when considering the probability of C2's words, it is reasonable to collapse words from *all* other speakers (C0,C1,C3,C4, and C5) into one stream A as shown in the figure. This makes available to C2 the rest of the meeting to potentially condition on, although it does not distinguish between different speakers.

Our model, Equation 1, is different from most language modeling systems since our models condition on both previous words and another potential factor A. Such a model is easily represented using a *factored language model* (FLM), an idea introduced in (Bilmes and Kirchhoff, 2003; Kirchhoff et al., 2003), and incorporated into the SRILM toolkit (Stolcke, 2002). Note that a form of cross-side modeling was used by BBN (Schwartz, 2004), where in a multi-pass speech recognition system the output of a first-pass from one speaker is used to prime words in the language model for the other speaker.

3 Initial Evaluation

We evaluate MSLMs on three corpora: Switchboard-I, Switchboard Eval-2003, and ICSI Meeting data. In Switchboard-I, 6.83% of the words are overlapped in time, where we define w_1 and w_2 as being overlapped if $s(w_1) \leq s(w_2) < e(w_1)$ or $s(w_2) \leq s(w_1) < e(w_2)$, where $s(\cdot)$ and $e(\cdot)$ are the starting and ending time of a word.

The ICSI Meeting Recorder corpus (Janin et al., 2003) consists of a number of meeting conversations with three or more participants. The data we employed has 32 conversations, 35,000 sentences and 307,000 total words, where 8.5% of the words were overlapped. As mentioned previously, we collapse the words from all other speakers into one stream A as a conditioning set for W. The data consists of all speakers taking their turn being W.

To be used in an FLM, the words in each stream need to be aligned at discrete time points. Clearly, a_t should not come from w_t's future. Therefore, for each w_t, we use the closest previous A word in the past for a_t such that $s(w_{t-1}) \leq s(a_t) < s(w_t)$. Therefore, each a_t is used only once and no constraints are placed on a_t's end time. This is reasonable since one can often predict a speaker's word after it starts but before it completes.

We score using the model $P(w_t|w_{t-1}, w_{t-2}, a_t)$.[1] Different back-off strategies, including different back-off paths as well as combination methods (Bilmes and Kirchhoff, 2003), were tried and here we present the best results. The backoff order (for Switchboard-I and Meeting) first dropped a_t, then w_{t-2}, w_{t-1}, ending with the uniform distribution. For Switchboard eval-2003, we used a generalized parallel backoff mechanism. In all cases, modified Kneser-Ney smoothing (Chen and Goodman, 1998) was used at all back-off points.

Results on Switchboard-I and the meeting data employed 5-fold cross-validation. Training data for Switchboard eval-2003 consisted of all of Switchboard-I. In Switchboard eval-2003, hand-transcribed time marks are unavailable, so A was available only at the beginning of utterances of W.[2] Results (mean perplexities and standard deviations) are listed in Table 1 (Switchboard-I and meeting) and the $|V|$ column in Table 3.

Table 1: Perplexities from MSLM on Switchboard-I (swbd-I) and ICSI Meeting data (mr)

data	trigram	four-gram	mslm	reduction
swbd-I	73.2±0.4	73.7±0.4	68.5±0.3	6.2%
mr	87.4±4.6	89.5±4.9	78.4±2.7	10.3%

[1] In all cases, end of sentence tokens, </s>, were not scored to avoid artificially small perplexities arising when $w_t = a_t =$ </s>, since $P(\text{</s>}|\text{</s>})$ yields a high probability value.

[2] Having time-marks, say, via a forced alignment would likely improve our results.

In Table 1, the first column shows data set names. The second and third columns show our best baseline trigram and four-gram perplexities, both of which used interpolation and modified Kneser-Ney at every back-off point. The trigram outperforms the four-gram. The fourth column shows the perplexity results with MSLMs and the last column shows the MSLM's relative perplexity reduction over the (better) trigram baseline. This positive reduction indicates that for both data sets, the utilization of additional information from other speakers can better predict the words of the current speaker. The improvement is larger in the highly conversational meeting setting since additional speakers, and thus more interruptions, occur.

3.1 Analysis

It is elucidating at this point to identify when and how A-words can help predict W-words. We thus computed the log-probability ratio of $P(w_t|w_{t-1}, w_{t-2}, a_t)$ and the trigram $P(w_t|w_{t-1}, w_{t-2})$ evaluated on all test set tuples of form $(w_{t-2}, w_{t-1}w_t, a_t)$. When this ratio is large and positive, conditioning on a_t significantly increases the probability of w_t in the context of w_{t-1} and w_{t-2}. The opposite is true when the ratio is large and negative. To ensure the significance of our results, we define "large" to mean at least $10^{1.5} \approx 32$, so that using a_t makes w_t at least 32 times more (or less) probable. We chose 32 in a data-driven fashion, to be well above any spurious probability differences due to smoothing of different models.

At the first word of a phrase spoken by W, there are a number of cases of A words that significantly increase the probability of a W word relative to the trigram alone. This includes (in roughly decreasing order of probability) echos (e.g., when A says "Friday", W repeats it), greetings/partings (e.g., a W greeting is likely to follow an A greeting), paraphrases (e.g., "crappy" followed by "ugly", or "Indiana" followed by "Purdue"), *is-a* relationships (e.g., A saying "corporation" followed by W saying "dell", A-"actor" followed by W-"Swayze", A-"name" followed by W-"Patricia", etc.), and word completions. On the other hand, some A contexts (e.g., laughter) significantly decrease the probability of many W words.

Within a W phrase, other patterns emerge. In particular, some A words significantly decrease the probability that W will finish a commonly-used phrase. For example, in a trigram alone, $p(\text{bigger}|\text{and}, \text{bigger})$, $p(\text{forth}|\text{and}, \text{back})$, and $p(\text{easy}|\text{and}, \text{quick})$, all have high probability. When also conditioning on A, some A words significantly decrease the probability of finishing such phrases. For example, we find that $p(\text{easy}|\text{and}, \text{quick}, \text{"uh-hmm"}) \ll p(\text{easy}|\text{and}, \text{quick})$. A similar phenomena occurs for other commonly used phrases, but only when A has uttered words such as "yeah", "good", "ok", "[laughter]", "huh", etc. While one possible explanation of this is just due to decreased counts, we found that for such phrases $p(w_t|w_{t-1}, w_{t-2}, a_t) \ll \min_{w_{t-3} \in S} p_4(w_t|w_{t-1}, w_{t-2}, w_{t-3})$ where p_4 is a four-gram, $S = \{w : C(w_t, w_{t-1}, w_{t-2}, w) > 0\}$, and C is the 4-gram word count function for the switchboard training and test sets. Therefore, our hypothesis is that when W is in the process of uttering a predictable phrase and A indicates she knows what W will say, it is improbable that W will complete that phrase.

The examples above came from Switchboard-I, but we found similar phenomena in the other corpora.

4 Conditional Probability Clustering

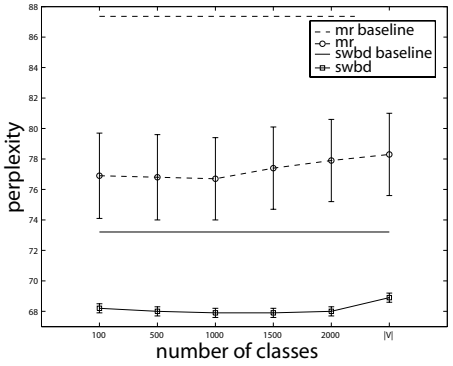

Figure 2: Class-based MSLM from MCMI clustering on Switchboard-I (swbd) and ICSI Meeting (mr) data.

Table 2: Three types of class-based MSLMs on Switchboard-I (swbd) and ICSI Meeting (mr) corpora

# of	swbd			mr		
classes	BROWN	MMI	MCMI	BROWN	MMI	MCMI
100	68.9±0.3	68.4±0.3	68.2±0.3	78.9±3.0	77.3±2.8	76.8±2.8
500	68.9±0.3	68.3±0.3	**67.9±0.3**	78.7±3.1	77.1±2.8	**76.7±2.8**
1000	68.9±0.3	68.2±0.3	67.9±0.3	79.0±3.1	77.2±2.7	76.9±2.8
1500	69.0±0.3	68.2±0.3	68.0±0.3	79.6±3.1	77.4±2.7	77.4±2.7
2000	69.0±0.3	68.3±0.3	68.0±0.3	80.1±3.1	77.6±2.7	77.9±2.7
$\|V\|$		68.5±0.3			78.3±2.7	

Table 3: Class-based MSLM on Switchboard Eval-2003

size	100	500	1000	1500	2000	$\|V\|$	3-gram	4-gram
ppl	65.8	65.5	65.6	65.7	66.1	67.9	72.1	76.3
% reduction	8.6	8.9	8.8	8.7	8.3	5.8	0	-5.8

Class-based language models (Brown et al., 1992; Whittaker and Woodland, 2003) yield great benefits when data sparseness abounds. SRILM (Stolcke, 2002) can produce classes to maximize the mutual information between the classes $I(C(w_t); C(w_{t-1}))$, as described in (Brown et al., 1992). More recently, a method for clustering words at different positions was developed (Yamamoto et al., 2001; Gao et al., 2002). Our goal is to produce classes that improve the scores $P(w_t|h_t) = P(w_t|w_{t-1}, w_{t-2}, C_1(a_t))$, what we call *class-based MSLMs*. In our case, the vocabulary for A is partitioned into classes by either maximizing conditional mutual information (MCMI) $I(w_t; C(a_t)|w_{t-1}, w_{t-2})$ or just

maximizing mutual information (MMI) $I(w_t; C(a_t))$. While such clusterings can perform poorly under low counts, our results show further consistent improvements.

Our new clustering procedures were implemented into the SRILM toolkit. When partitioned into smaller classes, the A-tokens are replaced by their corresponding class IDs. The result is then trained using the same factored language model as before. The resulting perplexities for the MCMI case are presented in Figure 2, where the horizontal axis shows the number of A-stream classes (the right-most shows the case before clustering), and the vertical axis shows average perplexity. In both data corpora, the average perplexities decrease after applying class-based MSLMs. For both Switchboard-I and the meeting data, the best result is achieved using 500 classes (7.1% and 12.2% improvements respectively).

To compare different clustering algorithms, results with the standard method of (Brown et al., 1992) (SRILM's `ngram-class`) are also reported. All the perplexities for these three types of class-based MSLMs are given in Table 2. For Switchboard-I, `ngram-class` does slightly better than without clustering. On the meeting data, it even does slightly worse than no clustering. Our MMI method does show a small improvement, and the perplexities are further (but not significantly) reduced using our MCMI method (but at the cost of much more computation during development).

We also show results on Switchboard eval-2003 in Table 3. We compare an optimized four-gram, a three-gram baseline, and various numbers of cluster sizes using our MCMI method and generalized backoff (Bilmes and Kirchhoff, 2003), which, (again) with 500 clusters, achieves an 8.9% relative improvement over the trigram.

5 Discussions and Conclusion

In this paper, novel multi-speaker language modeling (MSLM) is introduced and evaluated. After simply adding words from other speakers into a normal trigram context, the new model shows a reasonable improvement in perplexity. This model can be further improved when class-based cross-speaker information is employed. We also presented two different criteria for this clustering. The more complex criteria gives similar results to the simple one, presumably due to data sparseness. Even though Switchboard and meeting data are different in terms of topic, speaking style, and speaker number, one might more robustly learn cross-speaker information by training on the union of these two data sets.

There are a number of ways to extend this work. First, our current approach is purely data driven. One can imagine that higher level information (e.g., a dialog or other speech act) about the other speakers might be particularly important. Latent semantic analysis of stream A might also be usefully employed here. Furthermore, more than one word from stream A can be included in the context to provide additional predictive ability. With the meeting data, there may be a benefit to controlling for specific speakers based on their degree of influence. Alternatively, an MSLM might help identify the most influential speaker in a meeting by determining who most changes the probability of other speakers' words.

Moreover, the approach clearly suggests that a multi-speaker decoder in an automatic speech recognition (ASR) system might be beneficial. Once time marks for each word are provided in an N-best list, our MSLM technique can be used for rescoring. Additionally, such a decoder can easily be specified using graphical models (Bilmes and Zweig, 2002) in first-pass decodings.

We wish to thank Katrin Kirchhoff and the anonymous reviewers for useful comments on this work.

References

J. Bilmes and K. Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *Human Language Technology Conference*.

J. Bilmes and G. Zweig. 2002. The graphical models toolkit: An open source software system for speech and time-series processing. In *Proc. ICASSP*, June.

P. Brown, V. Della Pietra, P. deSouza, J. Lai, and R. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

S. Chen and J. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, August.

J. Gao, J. Goodman, G. Cao, and H. Li. 2002. Exploring asymmetric clustering for statistical langauge modeling. In *Proc. of ACL*, pages 183–190, July.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *Proc. ICASSP*, April.

K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, G. Ji, F. He, J. Henderson, D. Liu, M. Noamany, P. Schone, R. Schwartz, and D. Vergyri. 2003. Novel approaches to arabic speech recognition: Report from the 2002 Johns-Hopkins workshop. In *Proc. ICASSP*, April.

R. Schwartz. 2004. Personal communication.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int. Conf. on Spoken Language Processing*, September.

E. Whittaker and P. Woodland. 2003. Language modelling for Russian and English using words and classes. *Computer Speech and Language*, pages 87–104.

H. Yamamoto, S. Isogai, and Y. Sagisaka. 2001. Multi-class composite n-gram language model for spoken language processing using multiple word clusters. In *Proc. of ACL*, pages 531–538.