Speaker Adaptation of Language Models for Automatic Dialog Act Segmentation of Meetings

Jáchym Kolář[1], *Yang Liu*[2], *Elizabeth Shriberg*[3,4]

[1]University of West Bohemia, Department of Cybernetics, Pilsen, Czech Republic
[2]University of Texas at Dallas, Department of Computer Science, Richardson, TX, USA
[3]SRI International, USA [4]International Computer Science Institute, USA

jachym@kky.zcu.cz, yangl@hlt.utdallas.edu, ees@speech.sri.com

Abstract

Dialog act (DA) segmentation in meeting speech is important for meeting understanding. In this paper, we explore speaker adaptation of hidden event language models (LMs) for DA segmentation using the ICSI Meeting Corpus. Speaker adaptation is performed using a linear combination of the generic speaker-independent LM and an LM trained on only the data from individual speakers. We test the method on 20 frequent speakers, on both reference word transcripts and the output of automatic speech recognition. Results indicate improvements for 17 speakers on reference transcripts, and for 15 speakers on automatic transcripts. Overall, the speaker-adapted LM yields statistically significant improvement over the baseline LM for both test conditions.

Index Terms: language model adaptation, speaker adaptation, meetings, dialog act segmentation, sentence segmentation

1. Introduction

Segmentation of speech into sentence-like units is a crucial first step in spoken language processing, since many downstream language processing techniques (e.g., parsing, automatic summarization, information extraction and retrieval, machine translation) are typically trained on well-formatted input (such as written text). Several different approaches have been employed for sentence segmentation, including hidden Markov models (HMMs), multilayer perceptrons, maximum entropy, conditional random fields, AdaBoosting, and support vector machines [1, 2, 3, 4, 5, 6]. Many of these approaches rely on both acoustic (prosodic) and lexical information.

Studies on sentence segmentation have been conducted in different domains, including broadcast news, conversational telephone speech, and meetings. Our task in this paper is to automatically segment meeting recordings into sentences, or dialog acts (DAs) for this domain. In many real-life meeting applications, the speakers are often known beforehand and recorded on a separate channel. Moreover, many meetings have recurring participants, presenting the opportunity for adapting models to the individual talkers. Speaker adaptation methods were first successfully used in the cepstral domain for speech recognition [7, 8]. In [9], we evaluated speaker-dependent prosody models for DA segmentation in meetings. However, it was left unanswered whether speaker-dependent language models (LMs) would also benefit DA segmentation. In this paper we aim to address this question; our goal is to try to adapt the LM to capture speakers' idiosyncratic lexical patterns associated with DA boundaries.

General LM adaptation has been studied rather extensively in speech recognition and other language processing tasks, using both supervised [10, 11] and unsupervised [12, 13, 14] approaches. A useful survey of LM adaptation techniques is given in [15]. The typical approach is to use the test words to determine the topic for the test document, and then use the topic-specific LM or its combination with the generic LM. For the sentence segmentation task, [16] successfully used both acoustic and lexical data from another domain (with slightly different definition for sentence boundaries) to aid automatic sentence segmentation in meetings.

Although topic- and domain-based LM adaptation approaches have received significant attention in the literature, much less is known about LM adaptation for individual talkers. Akita and Kawahara [17] showed improved recognition performance using LM speaker adaptation by scaling the n-gram probabilities with the unigram probabilities estimated via probabilistic latent semantic analysis. Tur and Stolcke [18] demonstrated that unsupervised within-speaker LM adaptation significantly reduced word error rate in meeting recognition. Unlike previous work in DA segmentation (which typically focuses on features or modeling approaches in a speaker-independent fashion) or LM adaptation (mostly topic-based adaptation in the task of speech recognition), in this study, we investigate whether speaker adaptation of LMs may help in automatic DA segmentation of meetings.

The remainder of the paper is organized as follows. We describe our approach in Section 2. The experimental setup, results, and discussion are shown in Section 3; Section 4 provides conclusions.

2. Method

2.1. Hidden-event Language Model

For a given word sequence $w_1 w_2 ... w_i ... w_n$, the task of DA segmentation is to determine which inter-word boundaries correspond to a DA boundary. We label each inter-word boundary as either a within-unit boundary or a boundary between DAs. For example, in the utterance "yes we should be done by noon", there are two dialog acts: "yes" (an answer), and "we should be done by noon" (a statement). Each ends in a segmentation boundary.

We use a hidden event LM [19] to automatically detect

DA boundaries in the unstructured word sequence. The hidden event LM describes the joint distribution of words and DA boundaries, $P_{LM}(W, B)$. The model is trained by explicitly including the DA boundary as a token in the vocabulary in word-based n-gram LM. During testing, the hidden event LM uses the pair of word and DA boundary as the hidden state in HMM, and the words as observations, and performs a Viterbi or forward-backward decoding to find the DA boundaries given the word sequence. We used trigram LMs with modified Kneser-Ney smoothing [20] in the SRILM toolkit [21].

2.2. Speaker Adaptation Approach

To adapt the generic speaker-independent LM to a particular speaker, we use an interpolation approach. The speaker-adapted model SA is obtained from a linear combination of the speaker-independent model SI and a speaker-dependent model SD as follows:

$$P_{SA}(t_i|h_i; \lambda) = \lambda P_{SI}(t_i|h_i) + (1 - \lambda) P_{SD}(t_i|h_i) \qquad (1)$$

where t_i denotes a token (word or DA boundary) and h_i its history of $n - 1$ tokens in an n-gram LM. λ is a weighting factor that is empirically optimized on held-out data. We compare different approaches to estimate λs, as described in Section 3. Note that the SD data is already contained in the SI data for LM training; therefore, this interpolation does not help reduce out-of-vocabulary rate, it rather gives a larger weight to n-grams observed in the data corresponding to a particular speaker and is expected to be better suitable to this speaker.

3. Results and Discussion

3.1. Data and Experimental Setup

The ICSI meeting corpus [22] contains approximately 72 hours of multichannel conversational speech data and associated human transcripts, manually annotated for DAs [23]. We selected the top 20 speakers in terms of total words. Each speaker's data was split into a training set ($\sim$70% of data) and a test set ($\sim$30%), with the caveat that a speaker's recording in any particular meeting appeared in only one of the sets. Because of data sparsity, especially for the less frequent speakers, we did not use a separate development set, but rather jackknifed the test set in our experiments.

The total training set for speaker-independent models (comprising the training portions of the 20 analyzed speakers, as well as all data from 32 other less-frequent speakers) contains 567k words. Data set sizes for individual speakers are shown in Tables 1 and 2; size of training sets used for speaker adaptation (referred to as "adaptation sets") ranges from 5.2k to 115.2k words. We use the official corpus speaker IDs. The first letter of the ID denotes the sex of the speaker ("f" or "m"); the second letter indicates whether the speaker is a native ("e") or nonnative ("n") speaker of English.

We use two different test conditions: reference transcripts (REF) and speech recognition output (Speech-To-Text, STT). Recognition results were obtained using the state-of-the-art SRI CTS system [24], which was trained using no acoustic data or transcripts from the analyzed meeting corpus. To represent a fully automatic system, we also used automatic speech/nonspeech segmentation. Word error rates for this difficult data are still quite high; the STT system performed at 38.2% (on the whole corpus). To generate the "reference" DA boundaries for the STT words, we aligned the reference setup to the recognition output with the constraint that two aligned words could not occur further apart than a fixed time threshold.

A robust estimation of the interpolation weight λ in Eq (1) may be a problem because of data sparsity. In the jackknife approach, one half of speaker's test data is used to estimate λ for the other half, and vice versa. We test two methods for estimating λs. First, λs are estimated individually for each speaker. Second, we set λ as the average value of the interpolation weights across all the speakers. Note that in the latter approach, however, we still use only data from the first halves of individual test sets to estimate λ for the second halves, and vice versa. This approach eliminates having two significantly different values of λ for a single speaker, which did occur for some of the 20 analyzed speakers. It indicated that for those speakers, there is a mismatch in the two halves of the test data used in jackknife, and thus the weights were not optimized properly for the test set.

3.2. Evaluation Metrics

We measure DA segmentation performance using a "boundary error rate" [1]:

$$BER = \frac{I + M}{N_W} \quad [\%] \qquad (2)$$

where I denotes the number of false DA boundary insertions, M the number of misses, and N_W the number of words in the test set. In addition, we report overall results using another common metric, the NIST error rate. For DA segmentation, it is defined as the number of misclassified boundaries divided by the total number of DA boundaries in the reference. It can be expressed as

$$NIST = \frac{I + M}{N_{DA}} \quad [\%] \qquad (3)$$

where N_{DA} denotes the number of DA boundaries in the test set. Note that this error rate may be even higher than 100%.

3.3. Results for Individual Speakers

Table 1 shows a comparison of DA segmentation performance for the baseline speaker-independent LM and speaker-adapted LMs for individual speakers, using reference transcripts. The speakers displayed in the table are sorted according to the total numbers of words they have in the corpus. The numbers of words in adaptation and test sets are also listed. The results indicate that for 17 of 20 speakers, performance improved using both individual and global weights, and two other speakers improved only for one of the two interpolation methods. However, the degree of the improvement varies across particular speakers. For 8 talkers, the improvement was statistically significant at $p \leq 0.05$ using a Sign test for both methods. For 4 others it was significant for only one of the methods.

Table 2 reports the corresponding results for the STT conditions. Note that the test set sizes differ from the previous table, since the number of words in STT outputs is usually smaller than that in the corresponding reference. The results show that 15 speakers improved using both interpolation methods, while 4 other speakers improved just for one of the methods. Again, for 8 talkers, the improvement was significant at $p \leq 0.05$ for both methods, and for 4 others the improvement was significant

Table 1: *Boundary error rates (BER) [%] in REFerence conditions for individual speakers. ID=Speaker ID, #Adapt and #Test denote the number of words in the adaptation and test sets for each speaker, Baseline speaker-independent model performance, AdInW adaptation with individual weights, and AdGlW adaptation with global weights. IDs of speakers who improved using both methods are shown in boldface, * and ** indicate that the improvement is significant by a Sign test at $p <= 0.05$ for one or both methods, respectively.*

ID	#Adapt	#Test	Baseline	AdInW	AdGlW	ID	#Adapt	#Test	Baseline	AdInW	AdGlW
me013**	115.2k	51.2k	6.75	6.55	6.52	**mn052**	10.7k	3.8k	7.33	7.28	7.28
me011*	50.6k	24.8k	7.40	7.38	7.25	**mn021****	9.6k	4.1k	6.68	5.41	5.65
fe008**	50.6k	22.6k	7.51	7.12	7.16	**me003**	9.3k	3.6k	8.78	8.45	8.56
fe016*	32.0k	15.4k	7.35	7.22	7.18	**mn005****	7.7k	3.1k	7.83	7.01	6.92
mn015**	31.9k	14.7k	8.05	7.75	7.80	me045	8.1k	2.4k	8.90	8.94	8.90
me018*	31.8k	14.7k	6.64	6.43	6.45	**me025**	7.7k	2.4k	8.06	8.02	7.85
me010**	26.1k	12.6k	7.24	6.96	6.84	me006	6.9k	1.5k	9.53	10.32	9.47
mn007*	21.0k	10.1k	7.59	7.36	7.31	**me026**	5.2k	2.5k	5.80	5.76	5.80
mn017**	21.0k	7.1k	7.02	6.44	6.44	**me012****	5.3k	2.1k	6.85	6.29	6.29
mn082	13.3k	4.2k	6.33	6.28	6.21	fn002	5.9k	1.5k	10.92	10.79	11.33

Table 2: *Boundary error rates (BER) [%] in Speech-To-Text conditions for individual speakers. ID=Speaker ID, #Adapt and #Test denote the number of words in the adaptation and test sets for each speaker, Baseline speaker-independent model performance, AdInW adaptation with individual weights, and AdGlW adaptation with global weights. IDs of speakers who improved using both methods are shown in boldface, * and ** indicate that the improvement is significant by a Sign test at $p <= 0.05$ for one or both methods, respectively.*

ID	#Adapt	#Test	Baseline	AdInW	AdGlW	ID	#Adapt	#Test	Baseline	AdInW	AdGlW
me013**	115.2k	43.4k	8.29	8.16	8.18	**mn052***	10.7k	3.5k	10.87	10.49	10.17
me011**	50.6k	22.9k	8.81	8.59	8.51	**mn021***	9.6k	4.1k	8.20	7.83	7.73
fe008*	50.6k	19.5k	9.19	9.05	8.89	me003	9.3k	3.2k	9.36	9.36	9.33
fe016	32.0k	13.9k	8.42	8.40	8.31	**mn005****	7.7k	3.0k	11.47	8.94	10.42
mn015**	31.9k	13.7k	10.16	9.90	9.84	me045	8.1k	2.1k	11.08	11.42	11.27
me018**	31.8k	13.3k	8.12	7.83	7.90	me025	7.7k	1.6k	14.36	14.23	14.36
me010**	26.1k	11.3k	8.39	7.96	7.91	**me006***	6.9k	1.3k	10.94	10.01	10.40
mn007**	21.0k	8.4k	11.28	10.73	10.78	**me026**	5.2k	2.3k	7.35	7.01	6.88
mn017**	21.0k	6.0k	8.92	8.01	7.84	**me012**	5.3k	1.9k	8.68	8.15	8.31
mn082	13.3k	3.7k	10.37	10.45	10.10	fn002	5.9k	1.4k	13.40	13.40	12.89

for one method. An interesting observation is that for both testing conditions, the relative error reduction achieved by speaker adaptation is not correlated with the amount of adaptation data. This finding suggests that speakers differ inherently in how similar they are to the generic speaker-independent LM. Some talkers probably differ more and thus show more gain, even with less data.

3.4. Overall Results

An overall comparison of performance of baseline speaker-independent and speaker-adapted LMs is presented in Table 3. The test set contains 203k words for REF and 180k words for STT conditions. These results show that for both conditions, speaker-adapted LMs — with either global interpolation weights or individual weights — outperform the baseline. The overall improvements by LM speaker adaptation for both conditions are statistically significant at $p < 10^{-15}$, using a Sign test. Of the two weight options, global interpolation results in better performance; however, the difference between the two approaches is significant only at $p < 0.1$.

In speech-to-text conditions we also tried interpolating the speaker-independent model trained on reference transcriptions with a speaker-dependent model trained on the recognizer output. The idea was to allow the model also to adapt for error patterns typical for an individual talker. However, this adaptation performed less well than using reference transcriptions as the training data, which indicates that, at least with the amount of data available for our experiments, it is preferable to adapt LMs using clean data. In consequence it also suggests that prospective unsupervised approaches to LM speaker adaptation will perform less well than the supervised approach.

Table 3: *Overall boundary error rates (BER) [%] and NIST error rates [%] in REFerence and STT conditions*

Test conditions	REF		STT	
Metric	BER	NIST	BER	NIST
Baseline	7.30	48.56	9.06	68.65
Individual weights	7.02	46.74	8.79	66.61
Global weights	**6.99**	**46.56**	**8.76**	**66.38**
Adapt with STT data	N/A	N/A	8.97	68.03

4. Conclusions

We have explored speaker adaptation of hidden event language models for automatic DA segmentation of multiparty meetings. The speaker adaptation is based on a linear combination of the generic speaker-independent and speaker-dependent LMs. We evaluated the method on 20 frequent speakers with a wide range of total words available for LM adaptation. We examined the approach using both reference word transcripts and recognizer outputs. We found improvements for 17 speakers using reference transcripts, and for 15 speakers using automatic transcripts. Overall, we achieved a statistically significant improvement over the baseline LM for both testing conditions. The improvement was achieved even for some speakers who had only a relatively small amount of data available for adaptation. We conclude that speaker adaptation of LMs aids DA segmentation, and that future work should investigate the potential of speaker-specific modeling for other tasks. Other important areas for future extensions include the integration of lexical with prosodic or even multimodal information, and exploration of unsupervised approaches.

5. Acknowledgments

This material is based upon work supported by the Ministry of Education of the Czech Republic, projects No. 1M0567 and ME909, and by the U.S. Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

6. References

[1] Shriberg, E. et al.: “Prosody-based Automatic Segmentation of Speech into Sentences and Topics,” in *Speech Communication*, vol. 32, no. 1–2, pp. 127–154, 2000

[2] Warnke, V. et al.: “Integrated Dialog Act Segmentation and Classification Using Prosodic Features and Language Models” in *Proc. EUROSPEECH 97*, pp. 207–210, Rhodes, Greece, 1997

[3] Huang, J., Zweig, G.: “Maximum Entropy Model for Punctuation Annotation from Speech,” in *Proc. ICSLP 2002*, Denver, CO, 2002

[4] Liu, Y. et al.: “Using Conditional Random Fields for Sentence Boundary Detection in Speech,” in *Proc. ACL* , Ann Arbor, 2005

[5] Kolář, J., Shriberg, E., Liu, Y.: “Using Prosody for Automatic Sentence Segmentation of Multi-Party Meetings,” in *Text, Speech and Dialogue (TSD) 2006, Lecture Notes in Artificial Intelligence (LNAI)*, vol. 4188, pp. 629–636, Springer-Verlag, Berlin Heidelberg

[6] Akita, Y. et al.: “Sentence Boundary Detection of Spontaneous Japanese Using Statistical Language Model and Support Vector Machines,” in *Proc. Interspeech-ICSLP*, Pittsburgh, PA, 2006

[7] Gauvain, J.L., Lee, C.H.: “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Obsevation of Markov chains,” in *IEEE Transaction on Speech and Audio Processing*, vol.2, no.2., pp. 291–298, 1994

[8] Gales, M.J.F.: “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition,” in *Computer Speech and Language*, vol. 12., pp. 75–98, 1998

[9] Kolář, J., Shriberg, E., Liu, Y.: “On Speaker-Specific Prosodic Models for Automatic Dialog Act Segmentation of Multi-Party Meetings,” in *Proc. INTERSPEECH-ICSLP 2006*, Pittsburgh, PA, USA, 2006

[10] Kneser, R., Peters, J., Klakow, D.: “Language Model Adaptation Using Dynamic Marginals,” in *Proc. EUROSPEECH*, Rhodes, Greece, 1997

[11] Seymore, K., Rosenfeld, R.: “Using Story Topics for Language Model Adaptation,” in *Proc. Eurospeech 1997*, Rhodes, Greece, 1997

[12] Gretter, R., Riccardi, G.: “On-line Learning of Language Models with Word Error Probability Distributions,” in *Proc. ICASSP*, Salt Lake City, UT, 2001

[13] Niesler, T., Willett, D.: “Unsupervised Language Model Adaptation for Lecture Speech Transcription,” in *Proc. ICSLP*, Denver, CO, 2002

[14] Nanjo, H., Kawahara, T.: “Unsupervised Language Model Adaptation for Lecture Speech Recognition,” in *Proc.ISCA and IEEE SSPR*, Tokyo, Japan, 2003

[15] Bellegarda, J.R.: “Statistical Language Model Adaptation: Review and Perspectives,” in *Speech Communication*, vol. 42, pp. 93–108, 2004

[16] Cuendet, S., Hakkani-Tür, D., Tur, G.: “Model Adaptation for Sentence Segmentation from Speech,” in *Proc. IEEE/ACL Workshop on Spoken Language Technology (SLT)*, Aruba, 2006

[17] Akita, Y., Kawahara, T.: “Language Model Adaptation based on PLSA of Topics and Speakers,” in *Proc. INTERSPEECH-ICSLP 2004*, Jeju, Korea, 2004

[18] Tur, G., Stolcke, A.: “Unsupervised Language Model Adaptation for Meeting Recognition,” in *Proc. IEEE ICASSP 2007*, Honolulu, HI, 2007

[19] Stolcke, A. et al.: “Automatic Detection of Sentence Boundaries and Disfluencies Based on Recognized Words,” in *Proc. ICSLP*, pp. 2247–2250, Sydney, 1998

[20] Chen, S.F., Goodman, J.: “An Empirical Study of Smoothing Techniques for Language Modeling,” Tech.Rep. TR-10-98, CS Group, Harvard University, 1998

[21] Stolcke, A.: “SRILM - An Extensible Language Modeling Toolkit,” in *Proc. ICSLP*, Denver, CO, 2002

[22] Janin, A. et al.: “The ICSI Meeting Corpus,” in *Proc. ICASSP-2003*, Hong Kong, 2003

[23] Dhillon, R. et al.: “Meeting Recorder Project: Dialog Act Labeling Guide,” ICSI Technical Report TR-04-02, International Computer Science Institute, Berkeley, CA, 2004

[24] Zhu, Q. et al.: “Using MLP Features in SRI’s Conversational Speech Recognition System,” in *Proc. INTERSPEECH 2005*, pp. 2141–2144, Lisboa, 2005