Language Model Adaptation based on PLSA of Topics and Speakers

Yuya Akita[†‡] *Tatsuya Kawahara*[†‡]

[†] School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[‡] PRESTO, Japan Science and Technology Agency

Abstract

We address an adaptation method of statistical language models to topics and speaker characteristics for automatic transcription of meetings and discussions. A baseline language model is a mixture of two models, which are trained with different corpora covering various topics and speakers, respectively. Then, probabilistic latent semantic analysis (PLSA) is performed on the same respective corpora and the initial ASR result to provide unigram probabilities conditioned on input speech. Finally, the baseline model is adapted by scaling N-gram probabilities with these unigram probabilities. For speaker adaptation purpose, we make use of spontaneous speech corpus (CSJ) in which a large number of speakers gave talks for given topics. Experimental evaluation with real discussions showed that both topic and speaker adaptation improved test-set perplexity and word accuracy.

1. Introduction

Recent advances in computing and multimedia technologies have enabled digital audio archives of lectures, meetings and panel discussions to be constructed. Automatic speech recognition (ASR) plays an important role in producing these audio archives because transcription of speech is used to generate indices or summaries which are essential parts of an archive. We have been developing an automatic archiving system including ASR for panel discussions.

One significant problem in developing an ASR system dedicated to these kinds of speech is the difficulty of constructing a statistical language model matched to the target speech, because the amount of well-matched data is usually limited. So multiple text corpora covering different aspects of the target speech are combined. A typical implementation of this concept is interpolation of multiple language models covering various topics such as [1]. Adaptation based on interpolation can be performed by weighting or emphasizing models relevant to the input speech. Class-based language models[2] are also used for robust estimation of N-gram probabilities with limited or unmatched data. As a similar approach, several methods directly manipulating N-gram probabilities such as cache models[3] were proposed. With cache models, probabilities of N-grams involving the preceding context are increased.

With long and spontaneous speech, especially by multiple speakers (for example, in discussions), one other aspect related to speaker characteristics should be considered. Each speaker has his or her own speaking style, for instance, usage of fillers, end-of-sentence expressions and favorite phrases, and these linguistic phenomena are frequently observed in such speech. Individual speaking style cannot be handled properly by a uniform language model and should be modeled separately. Previous studies of language model adaptation have mainly focused on the characteristics of topics and have not surveyed the possibilities of speaker adaptation.

In this paper, we adapt a language model that takes speakers as well as topics into account. We adopt a probabilistic latent semantic analysis (PLSA) framework[4] and perform unigram scaling[5] based on PLSA to effectively operate N-gram probabilities. Topic and speaker characteristics are covered by different corpora, and PLSA is performed using each corpus. Scaling based on topic-based and speaker-based PLSA is combined, and a baseline language model is adapted. The proposed method is evaluated with real panel discussions.

2. Overview of Proposed Approach

Our approach aims at improving the prediction performance of a language model by taking into account both topic and speaker characteristics. Topic adaptation is performed for every discussion, because agendas of discussions differ every time, while the themes in a single session rarely change. On the other hand, speaker adaptation is performed for every speaker, since the style of speaking may be different for each speaker. In addition, unsupervised adaptation is required due to the limited amount of prior knowledge on each topic and participant.

As a framework of adaptation, we adopt PLSA. It is a kind of sub-space method in a probabilistic manner, and topics and speakers are characterized as dimensions in a sub-space. This sub-space is preliminarily constructed using large-scale text corpora. Using this sub-space, N-gram probabilities conditioned on the input text are directly estimated. Thus, unsupervised adaptation is conducted using a transcription by the baseline system to improve it by rescoring or redecoding.

PLSA is advantageous because it tries to fit an input text to the topic/speaker sub-spaces in a probabilistic manner rather than determining a topic/speaker or selecting texts. Especially for speaker characteristics that are relatively indistinct, PLSA is expected to be more suitable. Actually, we attempted interpolation based on text

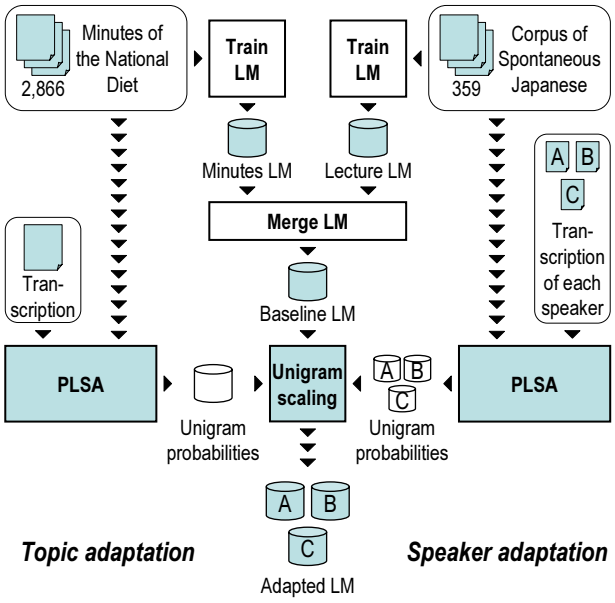

Figure 1: Flowchart of proposed method

selection by latent semantic analysis (LSA)[6], however, the resulting models could not improve prediction performance.

Figure 1 shows proposed adaptation method based on the PLSA framework. In advance, topic-oriented and speaker-oriented language models are trained using corpora covering variations of topics and speakers, respectively. Then, the baseline trigram language model is composed by merging these two models. The synthesis approach is a practical solution when sufficient size of the matched corpus is not available. The topic-oriented corpus can be easily collected from newspaper texts or Web pages. For the speaker-oriented corpus, we make use of the Corpus of Spontaneous Japanese (CSJ)[7], in which a large number (359) of speakers give talks on a couple of given topics such as hobby and travel. PLSA is performed separately for the topic-oriented and speaker-oriented corpora, and respective sub-spaces are constructed. For input speech, we conduct automatic speaker indexing and speech recognition to produce speaker labels and initial transcription. Language model adaptation is done by projecting the transcription into the sub-spaces. For speaker adaptation, this projection is done for transcription of each speaker. The projection is performed only for unigrams; bigram and trigram probabilities are approximately calculated using a scaling technique. Finally, an adapted language model is generated by interpolating the topic-projected and speaker-projected models.

3. Test-set Discussions

We compiled a corpus of panel discussions using a TV program "Sunday Discussion" broadcasted by NHK (the Japan Broadcasting Corporation). This program shows discussions on current political and economic topics by politicians, economists and experts in the fields. A chairperson also takes part and prompts the speakers. The duration of each discussion is one hour. The speech was segmented into utterances based on detection of short pauses longer than 400 milliseconds. Ten discussions of different themes are used for the experiments. The average number of utterances per session is 550.

Table 1: Specifications of language models

Model	Minutes	Lecture	Baseline
Training corpus	Minutes of the National Diet (1999-2002)	Corpus of Spontaneous Japanese	—
#Words	70M	2.9M	—
#Uniq. words	72K	37K	—
#Documents	2,866	359	—
Vocab. size	29K	5.8K	30K
Perplexity	187.50	111.89	105.62
OOV rate	4.78%	10.02%	2.13%

4. Training Corpora and Baseline Models

Two different corpora are used to cover topic and speaker characteristics in test discussions. The specifications of the corpora and derived language models are shown in Table 1.

The *Minutes* model is used to cover topics in the test discussions and trained from the minutes of the National Diet (Congress) of Japan for 1999 to 2002. Documents in the minutes are separated by the kind and date of meetings, and the total number of documents is 2,866. Utterances in these meetings are faithfully transcribed, but typical colloquial expressions such as fillers and end-of-sentence expressions are deleted or modified for documentation.

The *Lecture* model is used to cover speaker characteristics and trained from extemporaneous public speeches in the Corpus of Spontaneous Japanese (CSJ). Topics in these lectures are not relevant to those of the test discussions; typical examples are "an impressive experience in my life," "a journey to a foreign country" and so on. A single speaker talks on several topics, with content varying little from speaker to speaker. The total number of talks is 1,245. Then, a total of 359 documents are generated by concatenating those made by the same speaker, and these are used for PLSA. Consequently, speaker characteristics are expected to be mainly extracted by PLSA.

The baseline language model is constructed by linearly interpolating the *Minutes* and *Lecture* models. Their weights are preliminarily surveyed and determined to be 0.5 and 0.5, respectively. The test-set perplexity (PP) and Out-Of-Vocabulary (OOV) rates in Table 1 are computed using transcriptions of the test discussions. The PP and OOV rates are drastically improved by the combina-

tion and the effect of the *Lecture* model for spontaneous speech expressions is demonstrated because it does not contain topic words relevant to the test discussions.

5. PLSA of Topics and Speakers

PLSA is originally a method to characterize documents in a corpus. Each document is characterized in a probabilistic space, i.e., the coordinates of a document consist of word unigram probabilities. Then, a sub-space is constructed so that all documents are best discriminated. A new document is projected to the space in a probabilistic manner and word probabilities for this document d is estimated using

$$P(w|d) = \sum_{j=1}^{N} P(w|t_j)P(t_j|d), \quad (1)$$

where t_j is an unseen variable known as a latent variable, and N is the total number of latent variables (i.e., dimensions of the sub-space). Probabilities $\{P(w|t_j)\}$ and $\{P(t_j|d)\}$ correspond to the base of the sub-space and to the coordinates of document d in the sub-space, respectively. These probabilities are estimated using EM algorithm, and a detailed description is given in [4].

In the proposed method, PLSA is performed for topic-oriented and speaker-oriented corpora. For the former, all of the initial transcriptions by ASR are used, while PLSA is performed for every speaker for the latter. Speaker labels of input speech are obtained in a completely unsupervised manner, described in our previous work[8]. The average speaker indexing accuracy for the test discussions is 97%.

We made preliminary experiments to investigate the effect of a number of latent variables and determine optimal one. Figures 2 and 3 show the reduction rate of perplexity (PP) against the number of latent variables. In addition to automatic transcription (the "ASR result" in the Figures), manual transcription ("Transcription") was used as a reference. The reduction rate converged when the number of latent variables was around 200. Therefore, we determined that the operating points of latent variables were 250 and 200 for topic-based and speaker-based PLSA, respectively. Note that the reduction rate for the ASR result was smaller than that for manual transcription, since the ASR result contains errors.

6. Integrated Scaling of N-grams

For bigrams and trigrams, PLSA is approximately performed using a scaling technique, since reliable estimation of such a large number of N-gram entries is not easy. The estimation of trigram $w_{i-2}w_{i-1}w_i$ based on unigram scaling is formulated in Equation 2.

$$P'(w_i|w_{i-2}w_{i-1}) \propto \frac{P(w_i|d)}{P(w_i)} P(w_i|w_{i-2}w_{i-1}), \quad (2)$$

where $P(w_i)$ is a unigram probability in the baseline language model, and $P(w_i|d)$ is that obtained from PLSA.

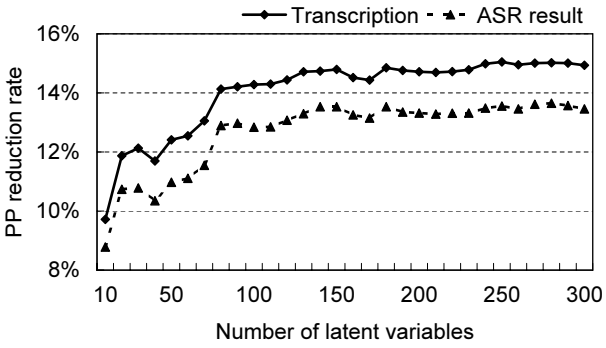

Figure 2: Effect of number of latent variables on PP reduction rate for the topic-oriented *Minutes* model

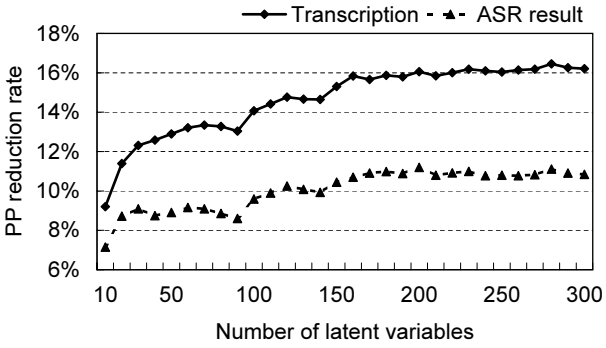

Figure 3: Effect of number of latent variables on PP reduction rate for the speaker-oriented *Lecture* model

In this work, unigram probabilities obtained by topic and speaker adaptation are integrated to scale the baseline language model. We use weighted sum of them:

$$P(w|d) = \lambda P_t(w|d) + (1-\lambda)P_s(w|d), \quad (3)$$

where P_t and P_s are unigram probabilities obtained by PLSA on topics and speakers, respectively. As the vocabulary in the two language models differs, the probability of words that never appear is set to 0. The value of interpolation weight λ is same as that used for the interpolation of the baseline language models described in Section 4.

7. Experimental Evaluation

We evaluated the proposed adaptation method using the test discussions. Perplexity (PP) and word accuracy are used to evaluate the performance. As for ASR, our decoder Julius[9] rev. 3.4.2 was used, and sequential decoding was performed. The acoustic model was triphone HMM trained from the CSJ. Unsupervised MLLR speaker adaptation was performed using speaker labels based on the unsupervised speaker indexing.

The reduction of PP with the proposed method using manual and automatic transcriptions is shown in Figures

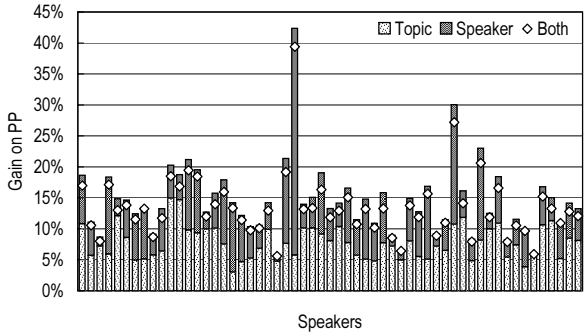

Figure 4: Improvement of PP for each speaker (PLSA for manual transcriptions)

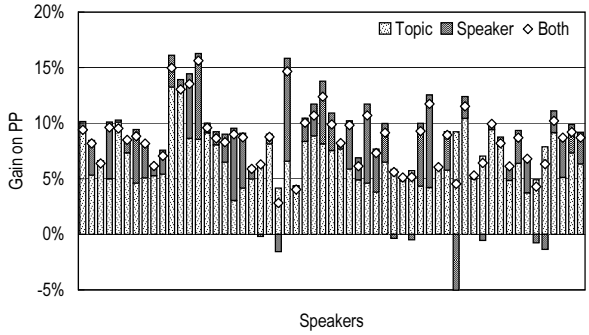

Figure 5: Improvement of PP for each speaker (PLSA for ASR results)

4 and 5, respectively. "Topic" and "Speaker" are reduction rates by only topic adaptation or speaker adaptation, respectively. "Both" is that with both adaptation methods. Average reduction rates of 7.78% and 6.82% were obtained by topic and speaker adaptation, respectively. The extent of the mis-match with the baseline language model is reflected in performance of adaptation, especially speaker adaptation, and caused larger differences in the reduction rate among speakers. The improvement by both adaptation was 13.41% and nearly equal to the sum of those by individual adaptation. This demonstrates that the weighted sum of unigram probabilities preserves the effectiveness of both adaptation methods.

Adaptation using automatic transcription showed slightly lower, but still significant effect as in Figure 5. Reduction rates of 6.66% and 2.30% were obtained by topic and speaker adaptation, respectively. Both adaptation achieved a 8.54% reduction rate.

Figure 6 shows the word accuracy of each discussion with an adapted language model based on PLSA using automatic transcription. The average accuracy for the baseline language model was 59.7%, and those for the topic-adapted, speaker-adapted and both-adapted models were 60.5%, 60.1% and 60.7%, respectively. The effect of the proposed method was confirmed by the ASR accuracy as well as perplexity.

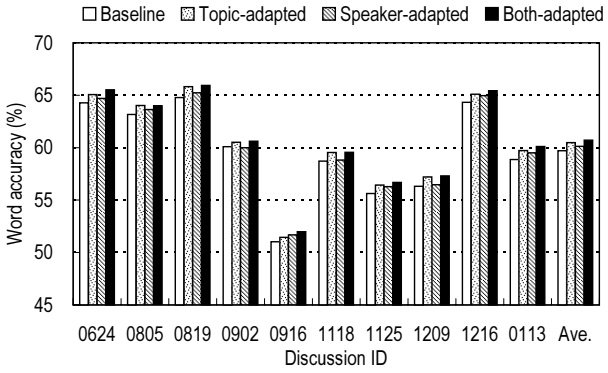

Figure 6: Word accuracy for each discussion (PLSA for ASR results)

8. Conclusion

We have presented a method of language model adaptation on topics and speakers using a PLSA framework. The baseline language model is constructed using two corpora covering topic and speaker characteristics. PLSA is performed using these respective corpora and the initial transcription, and then unigram probabilities adapted to the input transcription are obtained. N-gram entries in the baseline language model are scaled using these probabilities, and finally an adapted language model is generated. The proposed method showed improved performance with respect to both perplexity and word accuracy for real panel discussions.

9. References

[1] R. Kneser and V. Steinbiss. On the Dynamic Adaptation of Stochastic Language Models. In *Proc. ICASSP*, 1993.

[2] G. Moore and S. Young. Class-based Language Model Adaptation using Mixtures of Weights. In *Proc. ICSLP*, 2000.

[3] P. Clarkson and A. J. Robinson. Language Model Adaptation using Mixtures and an Exponentially Decaying Cache. In *Proc. ICASSP*, 1997.

[4] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. SIG-IR*, 1999.

[5] D. Gildea and T. Hofmann. Topic-based Language Models using EM. In *Proc. Eurospeech*, 1999.

[6] J. R. Bellegarda. Exploiting Latent Semantic Information in Statistical Language Modeling. *Proc. IEEE*, 88(2):1279–1296, 2000.

[7] S. Furui, K. Maekawa, and H. Isahara. Toward the Realization of Spontaneous Speech Recognition — Introduction of a Japanese Priority Program and Preliminary Results —. In *Proc. ICSLP*, 2000.

[8] Y. Akita and T. Kawahara. Unsupervised Speaker Indexing using Anchor Models and Automatic Transcription of Discussions. In *Proc. Eurospeech*, 2003.

[9] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano. Continuous Speech Recognition Consortium — An Open Repository for CSR Tools and Models —. In *Proc. IEEE Int'l Conf. on Language Resources and Evaluation*, 2002.