PLSA-based Topic Detection in Meetings for Adaptation of Lexicon and Language Model

Yuya Akita[†‡] *Yusuke Nemoto*[†] *Tatsuya Kawahara*[†‡]

[†] School of Informatics, Kyoto University,
[‡] Academic Center for Computing and Media Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

A topic detection approach based on a probabilistic framework is proposed to realize topic adaptation of speech recognition systems for long speech archives such as meetings. Since topics in such speech are not clearly defined unlike news stories, we adopt a probabilistic representation of topics based on probabilistic latent semantic analysis (PLSA). A topical sub-space is constructed by PLSA, and speech segments are projected to the subspace, then each segment is represented by a vector which consists of topic probabilities obtained by the projection. Topic detection is performed by clustering these vectors, and topic adaptation is done by collecting relevant texts based on the similarity in this probabilistic representation. In experimental evaluations, the proposed approach demonstrated significant reduction of perplexity and out-of-vocabulary rates as well as robustness against ASR errors.

Index Terms: Language model, lexicon, topic adaptation, topic detection, PLSA

1. Introduction

Recently, the major target of automatic speech recognition (ASR) research has been shifted to long and spontaneous speech like meetings [1] and parliamentary speeches [2, 3]. We also have been developing an ASR system for meetings in the National Diet of Japan[4]. One of major characteristics of such speech is a variety of topics, for example, people in parliamentary meetings discuss various current issues, in which new words frequently appear following new events in the society. For accurate and useful transcription of these meetings, a lexicon and language model of the ASR system should be updated periodically to cover the latest topics. On the other hand, the language model should be adapted to each topic to improve the prediction ability. This adaptation can usually be done by collecting relevant texts and interpolating with the general model. Apparently, there is a trade-off between the coverage of a variety of topics and the prediction ability for a particular topic, therefore, an optimal point for the reasonable coverage and prediction ability should be determined at the time of adaptation.

Meetings have no explicit topic boundaries, so they should be segmented for adaptation by detecting topics. In many of previous studies such as [5], however, topic adaptation is performed for preliminarily provided speech segments like broadcast news stories, and these studies did not consider how to determine topic boundaries in a speech. Automatic topic detection has been carried out mainly on newspaper corpora and broadcast news databases, for example, the topic detection and tracking (TDT) task [6]. In meetings, topics gradually change, so they are less distinct than those in the TDT task and more difficult to be segmented. Banerjee [7] applied a topic detection approach based on word occurrence statistics to manual transcription of meetings, however, the approach was not evaluated on ASR transcripts. Although some studies combine topic detection with ASR for spontaneous speech [8, 9], these studies were mainly intended for indexing purpose, and detection for topic adaptation has not been fully investigated.

In this paper, we propose topic adaptation combined with automatic topic detection for meeting speech. To deal with vagueness of topics in meetings, we introduce a probabilistic latent semantic analysis (PLSA) [10] to represent topic characteristics. Speech segments are clustered into topic segments using PLSA-based feature vectors, and adaptation of a lexicon and language model is performed for each segment by collecting relevant texts based on the PLSA-based similarity measure. In this work, the proposed method is evaluated with a meeting corpus of the National Diet of Japan.

2. PLSA for topic representation

In the proposed approach, topical features of speech are extracted by using the PLSA framework. PLSA is originally used to characterize documents in a corpus using word occurrence statistics. A topical sub-space, where each dimension represents some topic, is constructed by the expectation-maximization (EM) algorithm with a corpus of topically-segmented documents. The dimensions are optimally determined to distinguish documents in the training corpus. By projecting a document d into this subspace, a set of word occurrence probabilities $\{P(w|d)\}$ are obtained:

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

where $\{t_j\}$ are latent variables corresponding to dimensions, i.e., topics. N is a number of latent variables, i.e., dimensions of the sub-space.

In Equation (1), posterior probabilities $\{P(t_j|d)\}$ represent coordinates of the document d in the topical sub-space. These probabilities can be regarded as relevance measures for corresponding topics, therefore, we propose to use these probabilities as a topical feature vector v_d of the spoken document d.

$$v_d = \begin{pmatrix} P(t_1|d) \\ P(t_2|d) \\ \cdots \\ P(t_N|d) \end{pmatrix} \quad (2)$$

PLSA-based vectors have two advantages over conventional methods. In previous works, topical characteristics are often represented using vectors composed by word frequencies, which do not always provide good similarity measure, since they have a large number of dimensions. On the other hand, the number of dimensions of the proposed PLSA-based vectors is significantly smaller than that of the word frequency vectors, and each dimension has higher discriminant ability than the word frequencies. Also, word frequency vectors are sensitive to word errors caused by ASR, while PLSA-based vectors are expected to be more robust for errors because of the projection to a sub-space.

3. Proposed adaptation framework

Figure 1 shows an overview of the proposed method utilizing PLSA-based feature vectors. The proposed method consists of three stages.

In the first stage, topic detection is performed on an initial ASR transcript which is obtained by using the baseline lexicon and language model. An input transcript is preliminarily segmented into "speaker turns" when the speaker is changed. Consequently, utterances in each speaker turn are made by the same speaker. These speaker turns are individually projected into a PLSA-based topic sub-space, and a feature vector is derived for every turn. Then, by clustering these feature vectors, speaker turns are chunked into topic segments. Feature vectors are recalculated for each topic segment.

In the second stage, topically relevant text documents are collected from a large-scale text database such as a newspaper corpus. The same feature vector is extracted from each document in this database in advance, and these vectors are used for similarity calculation between each of these documents and topic segments obtained by the previous stage. Based on the similarity, relevant texts are selected for each topic segment.

In the final stage, the lexicon and language model used to obtain the initial ASR transcript are adapted to every topic segment using these selected texts. Out-of-vocabulary (OOV) words found in the selected texts are added to the initial lexicon. Also, a new language model is trained using the collected texts and the initial language model is linearly interpolated with this model.

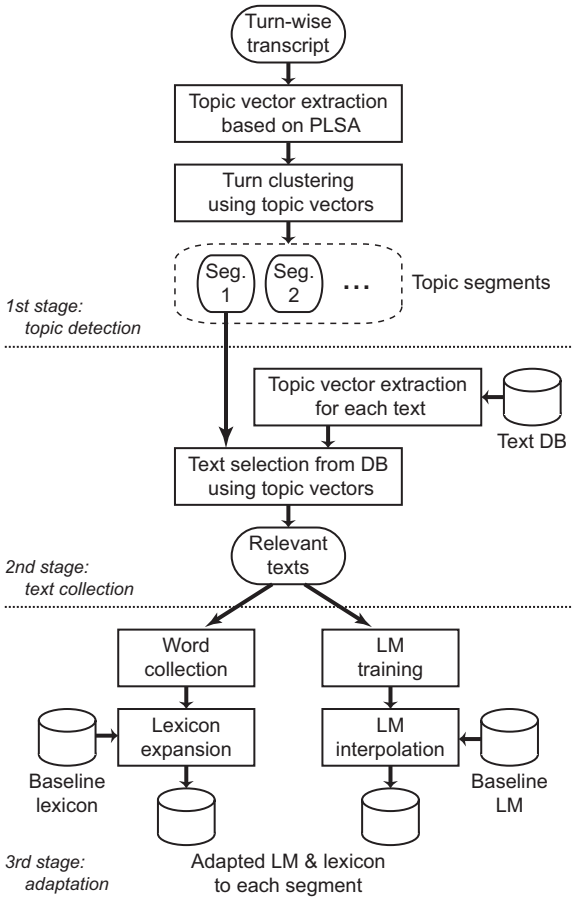

Figure 1: Flowchart of the proposed method

4. Topic detection using topical feature vectors

Each speaker turns are projected into the topic sub-space, and topic feature vectors are calculated. Speaker turns are then clustered based on the similarity between pairs of these vectors. We adopt an on-line method in which judgment of detection is sequentially performed for speaker turns in order of occurrence. If a similarity measure between new speaker turn d_j and current topic segment s_i is larger than a threshold θ_1, d_j is included in the segment s_i, otherwise a new topic segment s_{i+1} begins with d_j.

As the similarity measure between a topic segment s_i and a speaker segment d_j, we adopt the cosine distance:

$$sim_{cos}(s_i, d_j) = \frac{v_{s_i} \cdot v_{d_j}}{|v_{s_i}||v_{d_j}|} \quad (3)$$

$$v_{s_i} = \frac{1}{D_{s_i}} \sum_{d_k \in s_j} v_{d_k} \quad (4)$$

where D_{s_i} is a number of speaker turns in a topic segment s_i. In this work our primary concern is real-time processing for rapid generation of transcription, therefore, we adopt the on-line method and the cosine distance whose computation time is relatively smaller.

5. Adaptation of lexicon and language model

Next, relevant texts are selected from a document database for every topic segment, and expansion of a lexicon and interpolation of a language model are performed.

For selection of texts, we use the same similarity measure as that used in speaker turn clustering described in the previous section. Feature vectors are calculated for each topic segment of the input transcript and each document in the document database in the same way. Using these vectors, similarity is calculated for every pair of a topic segment and a document, and documents whose similarity score is larger than a threshold θ_2 are selected for the target topic segment.

Then, a lexicon for recognition of the target segment is expanded by adding OOV words found in the selected texts. Also, a small language model is trained with these texts, and an adapted language model is derived by linearly interpolating the baseline language model and this model.

6. Experimental evaluation

6.1. Setup

The proposed approach was evaluated using real meeting data. We prepared a test-set speech of a meeting held on February 2003 in the National Diet (Congress) of Japan. The duration of the speech is about 5.5 hours, the total number of words is 63K, the number of speakers is 23 and the total number of speaker turns is 296.

For the baseline language model, a trigram model was constructed using four-year (1999-2002) minutes of the Diet, and interpolated with a conversational model trained with the Corpus of Spontaneous Japanese (CSJ). The vocabulary size of the baseline model is 29,720. The test-set perplexity and OOV rate on the test-set by the baseline model are 61.9 and 0.47%, respectively. The word error rate by the baseline ASR system is 19.8%.

The topic sub-space was constructed with the same minutes as used for training of the baseline language model. The minutes were split by the dates and kinds of meetings, and the total number of minutes was 2,866. We determined the dimension of the sub-space to 250, which resulted in the optimal adaptation performance of language model in our previous work [11]. As a database for text selection, we used the Mainichi newspaper database. Articles written from July 2002 to the previous day of the test-set meeting were used. Categories of news articles were limited to those relevant to political and economic issues, and obviously irrelevant categories such as sports, home and life were excluded. The number of articles used in this experiment is 41,714, and they contain 11M words in total. Interpolation weights for the baseline model and a newspaper-based model were preliminarily investigated and determined as 0.85 and 0.15, respectively. These weights were used throughout all experiments.

6.2. Results

Figures 2–5 show reduction of perplexity and OOV rates by the proposed adaptation method using manual transcription and ASR results. Baseline methods which adapt to the entire speech or individual speaker turns were also performed, and their results are shown as "1-segment" and "296-segments," respectively. As for the proposed method, four different numbers of topic segments were conducted, by changing the value of the threshold θ_1. Although these numbers were different between manual transcription and ASR results, similar numbers were selected. For every condition, adaptation was performed with various text sizes determined by the text selection threshold θ_2.

In case of manual transcription shown in Figures 2 and 3, perplexity and OOV rates by adaptation to the entire speech ("1-segment") are largest among six conditions. It is because topic words and expressions were not supplied sufficiently for several topics in the speech. Smaller perplexity and OOV rates were obtained by adaptation to individual speaker turns ("296-segments") since topic characteristics of each segment were more distinct than "1-segment."

By comparing with these baseline results, even smaller perplexity and OOV rates were achieved in almost all conditions of the proposed approach. When sufficient numbers of topic segments such as 38 and 65 were obtained, the smallest perplexity and OOV rates were realized regardless of the size of the collected text. Perplexity converged to the almost same value in case of 25, 38, 65 and 296 topic segments, however, those by the proposed approach (38 and 65 topic segments) converged with the smaller amount of texts. It suggests that topic segments produced by the proposed approach had suitable granularity for adaptation, and enabled efficient text collection. When collecting texts from external resources, the text selection threshold (θ_2 in this work) must be chosen carefully, since too much texts include irrelevant ones and lead to increase of perplexity, as shown in [12, 13]. In this point of view, the proposed method provides a wider peak as shown in Figure 2, so it makes the choice easier. When perplexity converged, i.e., texts of around 400K words were used, reduction of perplexity was 9%. When 3K words were added to the baseline lexicon, the OOV rate was reduced by 36%.

As for evaluations on the ASR results, similar results were obtained as shown in Figures 4 and 5. Reduction of perplexity using 400K texts was 8%, and that of OOV rate by adding 3K words was 40%. In spite of the word error rate of 20%, reduction of perplexity and OOV rates were almost same as those on manual transcription. Thus, the proposed method is robust against word errors.

7. Conclusions

We have proposed a PLSA-based topic detection approach dedicated to adaptation of a lexicon and language model. Feature vectors are generated by projecting speech segments to a PLSA-based topical sub-space, and

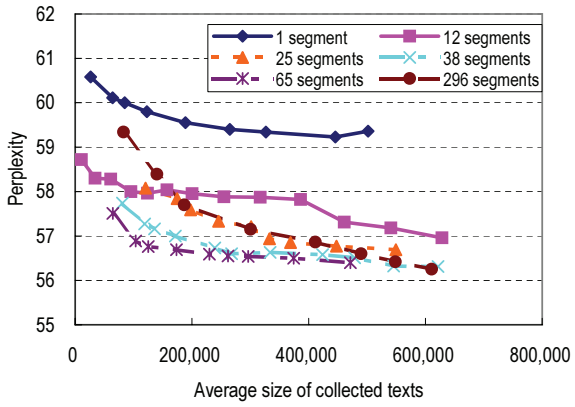

Figure 2: Perplexity with manual transcription

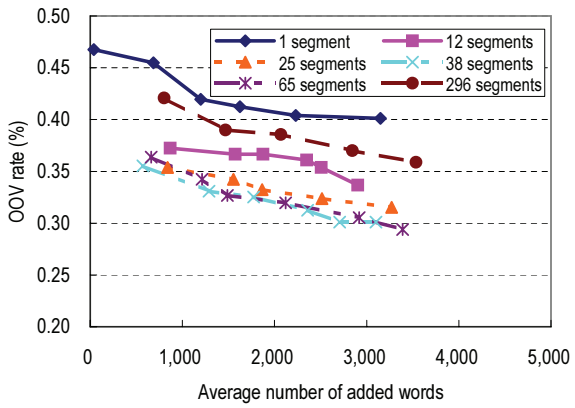

Figure 3: OOV rates with manual transcription

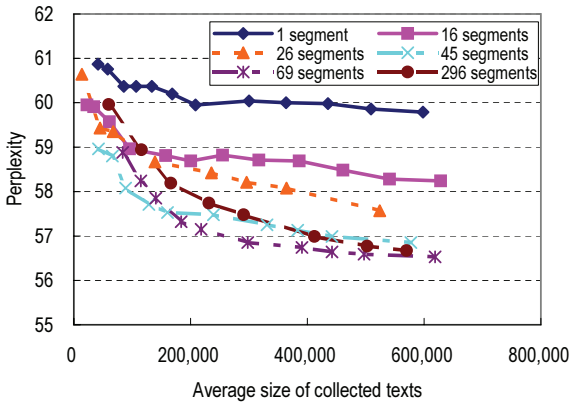

Figure 4: Perplexity with ASR results

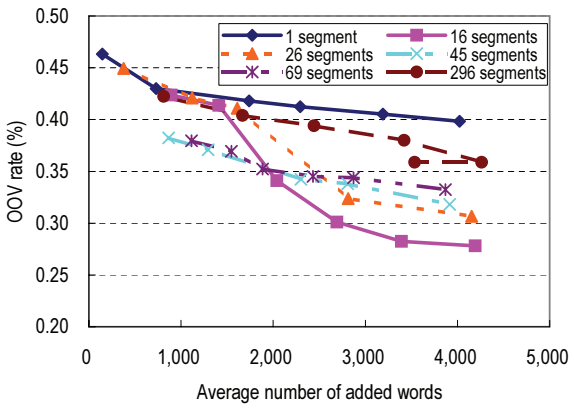

Figure 5: OOV rates with ASR results

topic segments are determined by clustering these vectors. Relevant texts are collected for each topic segment, then a lexicon and a language model are adapted using these texts. The proposed approach realized significant reduction of perplexity and OOV rates on experimental evaluation, and also demonstrated robustness for ASR errors.

8. Acknowledgment

This work is partly supported by Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communications, Japan.

9. References

[1] F. Metze, C. Fügen, Y. Pan, and A. Waibel, "Automatically Transcribing Meetings using Distant Microphones," in *Proc. ICASSP*, 2005.

[2] J. Loof, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schluter, and H. Ney, "The 2006 RWTH Parliamentary Speeches Transcription System," in *Proc. Interspeech*, 2006.

[3] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro, "The IBM 2006 Speech Transcription System for European Parliamentary Speeches," in *Proc. Interspeech*, 2006.

[4] Y. Akita and T. Kawahara, "Topic-independent Speaking-style Transformation of Language Model for Spontaneous Speech Recognition," in *Proc. ICASSP*, 2007.

[5] K. Ohtsuki, N. Hiroshima, M. Oku, and A. Imamura, "Unsupervised Vocabulary Expansion for Automatic Transcription of Broadcast News," in *Proc. ICASSP*, 2005.

[6] C. Cieri, D. Graff, S. Strassel, and N. Martey, "The TDT-3 Text and Speech Corpus," in *Proc. Topic Detection and Tracking Workshop*, 2000.

[7] S. Banerjee and A.I. Rudnicky, "A TextTiling Based Approach to Topic Boundary Detection in Meetings," in *Proc. Interspeech*, 2006.

[8] M. Franz, B. Ramabhadran, T. Ward, and M. Picheny, "Automated Transcription and Topic Segmentation of Large Spoken Archives," in *Proc. Eurospeech*, 2003.

[9] K. Bessho, K. Ohtsuki, N. Hiroshima, S. Matsunaga, and Y. Hayashi, "Topic Structure Extraction for Meeting Indexing," in *Proc. ICSLP*, 2004.

[10] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. SIG-IR*, 1999.

[11] Y. Akita and T. Kawahara, "Language Model Adaptation based on PLSA of Topics and Speakers," in *Proc. ICSLP*, 2004.

[12] R. Sarikaya, A. Gravano, and Y. Gao, "Rapid Language Model Development Using External Resources for New Spoken Dialog Domains," in *Proc. ICASSP*, 2005.

[13] T. Misu and T. Kawahara, "A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts," in *Proc. Interspeech*, 2006.