Language Modeling using PLSA-Based Topic HMM

Atsushi SAKO[1], *Tetsuya TAKIGUCHI*[2], *Yasuo ARIKI*[2]

[1]Department of Informatics and Electronics
[2]Department of Computer and System Engineering
Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN
sakoats@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

In this paper, we propose a PLSA-based language model for sports live speech. This model is implemented in unigram rescaling technique that combines a topic model and an n-gram. In conventional method, unigram rescaling is performed with a topic distribution estimated from a history of recognized transcription. This method can improve the performance; however it cannot express topic transition. Incorporating concept of topic transition, it is expected to improve the recognition performance. Thus the proposed method employs a "Topic HMM" instead of a history to estimate the topic distribution. The Topic HMM is a Discrete Ergodic HMM that expresses typical topic distributions and topic transition probabilities. Word accuracy results indicate an improvement over tri-gram and PLSA-based conventional method using a recognized history.

Index Terms: language modeling, text model, PLSA, HMM, speech recognition

1. Introduction

Recently large quantities of multimedia contents are broadcast and accessed through digital TV and WWW. In order to retrieve exactly what we want to know from them, automatic extraction of meta-information or structuring is strongly required. Sophisticated automatic speech recognition (ASR) plays an important role for extracting this kind of information because accurate transcription is inevitable. The purpose of this study is to improve the speech recognition accuracy for automatically transcribing sports live speech especially baseball commentary speech, in order to produce the closed caption and to structure the sports games for highlight scene retrieval.

As the sports live speech, we used radio speech instead of TV speech because it has much more information. However the radio speech is rather fast and noisy. Furthermore, it is disfluent due to rephrasing, repetition, mistake and grammatical deviation caused by spontaneous speaking style. To solve these problems, we proposed the adaptation techniques for acoustic model and language model [1] and the situation based language model [2].

In order to further improve the speech recognition accuracy, we focus on topic-based language models in this paper. Several topic-based language models have been studied; stochastic switching language model [3], Latent Semantic Analysis (LSA) based language model [4] or a PLSA-based language model using unigram rescaling technique [5]. SS N-gram requires large quantity of corpus however it is difficult to create large corpus in sports tasks. PLSA is a probabilistic model of LSA and a more compatible with a N-gram than LSA. Thus, in this paper, we focus on especially PLSA-based models.

The conventional PLSA-based model estimates a topic distribution using a "History" of recognized transcription. However, it cannot express topic transition. Considering topic transition, the recognition accuracy is improved because it enables to use proper language model for each topic. Consequently, we propose a new language model based on PLSA. The model expresses typical distributions of topics and transition probabilities between topics. We implemented this model as a Discrete Ergodic HMM which has discrete distribution in each state and transition probabilities between states. We call the HMM "Topic-HMM". Unigram probabilities are obtained from a distribution of a state through the algorithm described in Section 2. Moreover, tri-gram probabilities are also obtained from unigram rescaling technique. For each state of the Topic HMM, tri-gram is computed as a topic dependent language model. The experimental results show that the Topic HMM improves the performance of the word accuracy.

2. PLSA-Based Language Modeling

Probabilistic Latent Semantic Analysis (PLSA) [6] is a topic decomposition method for documents in a corpus. It is used to analyze topic distributions of documents and unigram distributions in a topic. The model is estimated from the co-occurrence probability of words and documents. Let d denote a document from a text corpus, w denote a word, and z denote a latent variable that represents a topic. Under the assumption that a document and a word are independent of each other given a latent variable, the conditional probability of generating a word from a document is

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d). \quad (1)$$

The $P(w|z)$ parameter is a unigram probability conditioned on a latent variable. The $P(z|d)$ parameter is a topic probability over each document. Note that, to distinguish a latent topic of PLSA from a topic of Topic-HMM, we call a topic of PLSA "a topic" and a topic of Topic-HMM "a state". Thus, the Topic HMM treats *states* as actual topics that consist of *topics* as latent topics. Additionally, *topic distribution* is defined as a vector consisting of topics. Namely, topic distribution is $(P(z_1|d), \cdots, P(z_K|d))^T$ for document d, where K is the number of latent topics.

Each parameter is estimated by the Expectation Maximization (EM) algorithm. The E-step is

$$P(z|d,w) = \frac{P(w|z)P(z|d)}{\sum_{z' \in Z} P(w|z')P(z'|d)}, \quad (2)$$

and the M-step is

$$P(w|z) = \frac{\sum_{d\in D} N(d,w)P(z|d,w)}{\sum_{w\in W}\sum_{d\in D} N(d,w)P(z|d,w)}, \quad (3)$$

$$P(z|d) = \frac{\sum_{w\in W} N(d,w)P(z|d,w)}{N(d)} \quad (4)$$

where N(d,w) is the number of the co-occurrences of the word w and the document d.

To use this PLSA-based model as a language model, it is proposed in [5] to compute the probabilities of words given histories $P(w|h)$. Hence, $P(w|h)$ is approximately computed as follows:

$$P(w|h_i) = \sum_z P(w|z)P(z|h_i), \quad (5)$$

$$P(z|h_i) = \frac{1}{i+1}\frac{P(w_i|z)P(z|h_{i-1})}{\sum_{z'} P(w_i|z')P(z'|h_{i-1})} + \frac{i}{i+1}P(z|h_{i-1}), \quad (6)$$

$$P(z|h_1) = P(z) = \frac{\sum_{w,d} N(d,w)P(z|d)}{\sum_{w,d} N(w,d)}. \quad (7)$$

However, the $P(w|h)$ parameter is only a mixture of unigram distribution. Thus, the *unigram rescaling* technique is proposed in [5], which combines the PLSA-based model with an n-gram as follow:

$$P(w_i|w_{i-1}w_{i-2}) \propto \frac{P(w_i|h_i)}{P(w_i)}P(w_i|w_{i-1}w_{i-2}). \quad (8)$$

3. Language modeling using Topic HMM

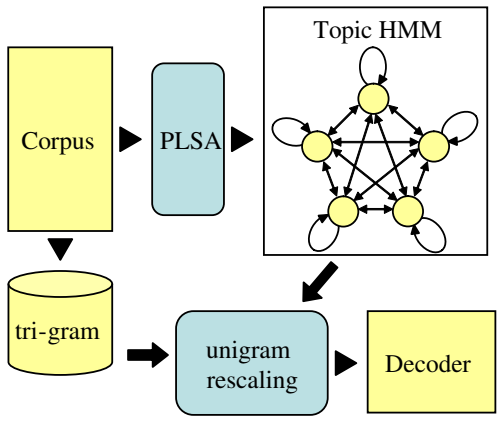

Figure 1: *Overview of proposed method.* The Topic HMM is learned using feature vectors obtained from topic distribution of each utterance. A state of HMM corresponds to a discrete distribution of topics. The decoding is performed with language models constructed by combining the Topic HMM and a tri-gram using unigram rescaling technique.

In this section, we describe how to construct a PLSA-based Topic HMM. Figure 1 shows an overview of the proposed method. First, PLSA is performed to estimate topic distribution $P(z|d)$ for all documents in a corpus and unigram distribution $P(w|z)$. Note that we employed an utterance as a document. There are about 8,000 utterances that consist of 5 to 20 words. Topic of each document is expressed as a vector consisting of probabilities $P(z_1|d)\cdots P(z_K|d)$ where K is the number of topics of PLSA.

Next, a Discrete Ergodic HMM (shown in figure 2) is trained using feature vectors obtained from PLSA. Initial distribution of each state is computed by K-means method. After Baum-Welch training, each state is a cluster that is collected from similar situations or topics. The mean vector of each state corresponds to the typical probabilities of topics. However, there is no guarantee that sum of mean vector elements becomes 1. Hence, the normalization is performed to be $\sum_{z\in Z} P(z|s) = 1$. A state transition probability of an Ergodic HMM is a topic transition probability.

Decoding is performed by driving the topic model described above. Let $\mathbf{W}$ be a word sequence and $\mathbf{S}$ be a state sequence of a Topic HMM. A language model is formulated as follows:

$$\begin{aligned} P(\mathbf{W}) &= \sum_{\mathbf{S}} P(\mathbf{W},\mathbf{S}) \\ &= \sum_{\mathbf{S}}\prod_i P(s_i|s_1^{i-1},w_1^{i-1})P(w_i|w_1^{i-1},s_1^i) \\ &\approx \max_S \prod_i P(s_i|s_{i-1})P(w_i|w_{i-2}^{i-1},s_i). \end{aligned} \quad (9)$$

Note that, in this research, a state transition probability $P(s_i|s_{i-1})$ is given only between utterances. This is because the Topic HMM is trained using utterances as a unit. To adjust the effect of a transition probability $P(s_i|s_{i-1})$ of the Topic HMM, scaling factor α is employed. Eq. 10 is derived from Eq. 9 with the scaling factor α:

$$P(\mathbf{W}) = \max_S \prod_i P(s_i|s_{i-1})^\alpha P(w_i|w_{i-2}^{i-1},s_i). \quad (10)$$

Here, Eq. 11 can be derived from Eq. 8 and Eq. 10,

$$P(w_i|w_{i-1}^{i-2},s_i) \propto \frac{P(w_i|s_i)}{P(w_i)}P(w_i|w_{i-1}^{i-2}). \quad (11)$$

Here, $P(w_i|s_i)$ means the word unigram probability of the state s_i of the Topic HMM. It is computed by Eq. 1. $P(z|s_i)$ corresponds to a component of a mean vector of the state.

Here, we describe how to recognize speeches using the proposed language model in detail. Figure 3 shows a process of decoding. Initially, the decoder knows $P(z_t|s_i)$, $P(w_i|s_i)$, $P(s_i|s_{i-1})$, $P(w_i)$ and $P(w_i|w_{i-1})$ because these are learned by PLSA and obtained from tri-gram probabilities. A language model for each state is constructed in the following steps. First, $P(w_i|s_i)$ is computed by Eq. 1 using a mean vector μ_{s_i} of a state of Topic-HMM.. Then, $P(w_i|w_i,s_i)$ is computed by Eq. 8 for each state, which is the language model for a state. For each utterance, speech recognition is performed, and then, the most likely sequence of states S is obtained by dynamic programming. Finally, the speech recognition result is the word sequence corresponding to the sequence of states.

4. Experiments

To evaluate the language model using a PLSA-based Topic HMM, speech recognition was performed. The test set is a commentary speech on a baseball live game. We used a commentary speech on a radio instead of a TV since it contains much more information. We performed the experiments using three methods; tri-gram, unigram rescaling from a recognized history (called "History"), and proposed method. In the next section, we describe the experimental conditions.

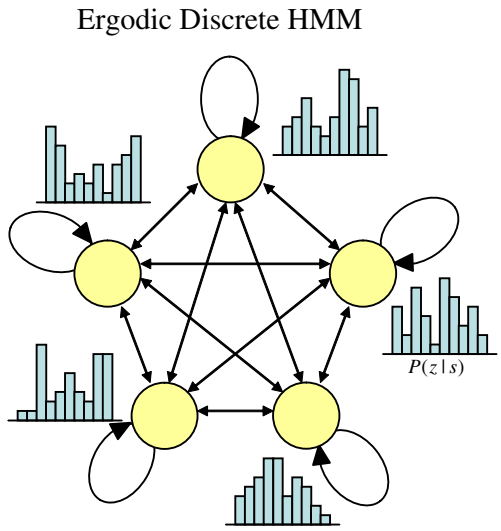

Figure 2: *A Discrete Ergodic HMM as a Topic HMM*. Each state represents an actual topic that consists of a mixture of topics.

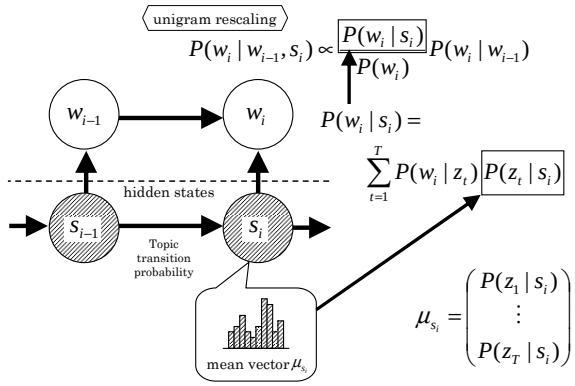

Figure 3: A decoding process.

4.1. Experimental Conditions

The acoustic model was syllable-based monophone HMM (left to right) [8]. The experimental conditions of acoustic model are summarized in table 1. The training data consisted of about 200,000 Japanese sentences (200 hours) spoken by 200 males in Corpus of Spontaneous Japanese (CSJ) [7]. Supervised acoustic model adaptation was performed using MLLR+MAP [9] with about 2 hours data that matched the test set.

The baseline of language model was a tri-gram language model. The training data consisted of 80,000 words collected from manual transcription of commentary speech on baseball games. The number of unique words is about 3,000. The domain and the vocabulary of the training data were matched with the test set.

Experiments were performed using decoder Julius [10] rev. 3.5. Because the performance depends on the number of topics or states, we performed experiments over various parameters.

Table 1: *Experimental conditions of acoustic model*

Sampling rate/Quantization	16 kHz / 16 bit
Feature vector	25 - order MFCC
Window	Hamming
Frame size/shift	20/10ms
# of phoneme categories	244 syllable
# of mixtures	32
# of states (Vowel)	5 states and 3 loops
# of states (Consonant+Vowel)	7 states and 5 loops

4.2. Experimental Results

The experimental results are summarized in table 2. These results are the best performances of each method. The accuracy for the "tri-gram" was 66.5%, for the "History" was 67.1% with 50 topics of PLSA, and for the "proposed method" was 69.9% with 70 topics of PLSA and 30 states of the Topic HMM. The Topic HMM improved the speech recognition performance. For example, there was an utterance such as "pitch and KARABURI (means strike out)", but in the conventional method, it was recognized as "pitch and TAMURA-RIN (name of player)" due to similar pronunciation. In the proposed method, it recognized correctly because the Topic HMM indicated that the probability of the player name after pitching is low.

Figure 4 shows the word accuracy of the proposed method over the number of topics of PLSA from 5 to 70 and the number of states of the Topic HMM from 5 to 60. The best performance was 69.9% with 70 topics and 30 states, and even the worst performance achieved 67.8%. However, it took much time to seek a topology with the best performance.

Figure 5 shows the effect of transition probability of the Topic HMM. In a case that the scaling factor is zero, the transition probability is not used. Here, we picked up typical three cases with the same level of accuracy. The numbers of the states of the Topic HMM for these cases are 15, 30 and 60, respectively. We can see the improvement of the word accuracy using transition probability in all cases.

Table 2: *Experimental results*

	tri-gram	History	Topic HMM
Word Accuracy	66.5%	67.1%	**69.9%**
Improvement	-	+0.6%	**+3.4%**

5. Conclusions

In this paper, we propose the language modeling method using a Topic HMM based on PLSA framework. The Topic HMM is learned from latent class distributions of each document estimated by PLSA. The decoding is performed using unigram rescaling technique to combine the Topic HMM and tri-gram models. We performed the experiments on the task of a commentary speech on a baseball game. The experimental results show that Topic-HMM improves the performance of the word accuracy.

In the future, we will perform experiments with general corpus such as CSJ or JNAS. Moreover, we will study automatic determination techniques of the Topic HMM topology such as the number of topics of PLSA and the number of states of the Topic HMM.

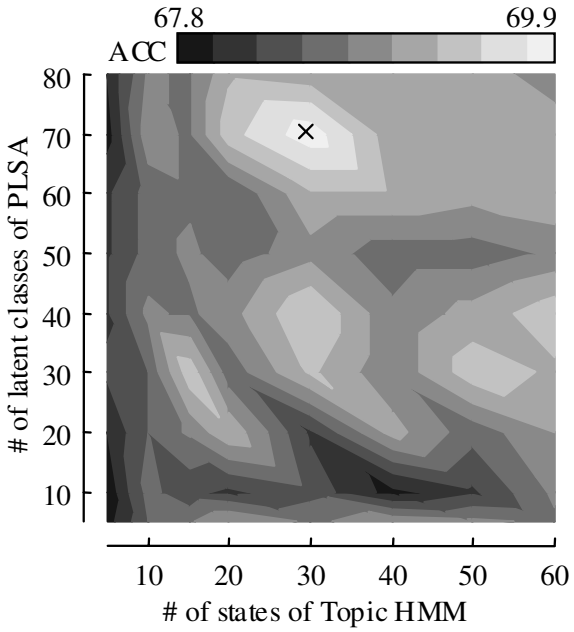

Figure 4: *Word accuracy results.* Vertical axis shows the number of topics of PLSA. Horizontal axis shows the number of states of the Topic HMM. The best performance was obtained with 70 classes and 30 states of PLSA and the Topic HMM, respectively.

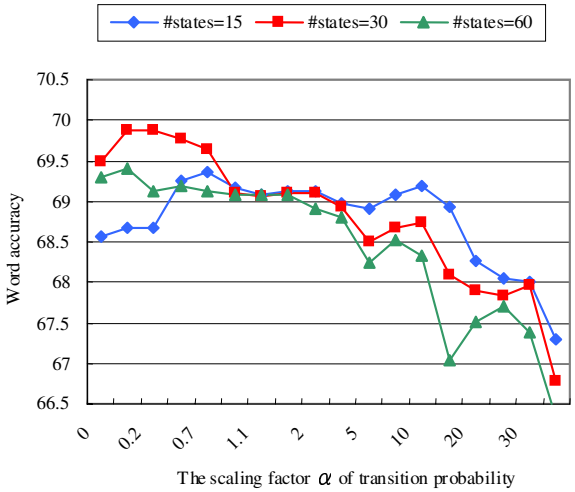

Figure 5: *Effect of transition probability of the Topic HMM on word accuracy.* Horizontal axis shows the scaling factor α of transition probability.

6. References

[1] Y. Ariki, T. Shigemori, T. Kaneko, J. Ogata and M. Fujimoto, "Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model", in Eurospeech2003, pp.1453-1456, 2003.

[2] A. Sako, Y. Ariki: "Structuring Baseball Live Games Based on Speech Recognition Using Task Dependent Knowledge and Emotion State Recognition", in ICASSP 2005, pp. 1049-1052, 2005.

[3] T. Nagano, M. Suzuki, A. Ito and S. Makino, "Language Modeling using Stochastic Switching N-gram", in Proceedings of the 18th International Congress on Acoustics, V, pp.3697-3700, 2004.

[4] Bellegarda, Jerome R., "Exploiting Latent Semantic Information in Statistical Language Modeling", in Proc. of IEEE, Vol. 88, Num. 8, pp. 1279-1296, 2000.

[5] D. Gildea and T. Hofmann, "Topic-Based Language Models Using EM", in Eurospeech'99, pp.2167-2170, 1999.

[6] T. Hofmann, "Probabilistic Latent Semantic Analysis", in Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99), 1999.

[7] S. Furui, K. Maekawa, H. Isahara, "Spontaneous Speech: Corpus and Processing Technology", The Corpus of Spontaneous Japanese, pp.1-6, 2002.

[8] J. Ogata and Y. Ariki, "Syllable-Based Acoustic Modeling for Japanese Spontaneous Speech Recognition", in Eurospeech 2003, pp.2513-2516, 2003-09.

[9] E. Thelen, X. Aubert and P. Beyerlein, "Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition", in ICASSP 1997, pp1035-1038, 1997.

[10] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Ito 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.