Unsupervised Segmentation of Phoneme Sequences based on Pitman-Yor Semi-Markov Model using Phoneme Length Context

Ryu Takeda and **Kazunori Komatani**
The Institute of Scientific and Industrial Research, Osaka University
8-1, Mihogaoka, Ibaraki, Osaka 567-0047, Japan
{rtakeda, komatani}@sanken.osaka-u.ac.jp

Abstract

Unsupervised segmentation of phoneme sequences is an essential process to obtain unknown words during spoken dialogues. In this segmentation, an input phoneme sequence without delimiters is converted into segmented sub-sequences corresponding to words. The Pitman-Yor semi-Markov model (PYSMM) is promising for this problem, but its performance degrades when it is applied to phoneme-level word segmentation. This is because of insufficient cues for the segmentation, e.g., homophones are improperly treated as single entries and their different contexts are also confused. We propose a phoneme-length context model for PYSMM to give a helpful cue at the phoneme-level and to predict succeeding segments more accurately. Our experiments showed that the peak performance with our context model outperformed those without such a context model by 0.045 at most in terms of F-measures of estimated segmentation.

1 Introduction

1.1 Motivation

The final goal of our current project is to achieve the development of robots or systems that acquire knowledge during spoken interactions between them and human beings in the open world. Unknown or new words appear frequently in our daily lives, and because their meanings may be different for the systems deployed in different areas, automatic lexicon acquisition is a useful function for maintenance-free spoken dialogue systems.

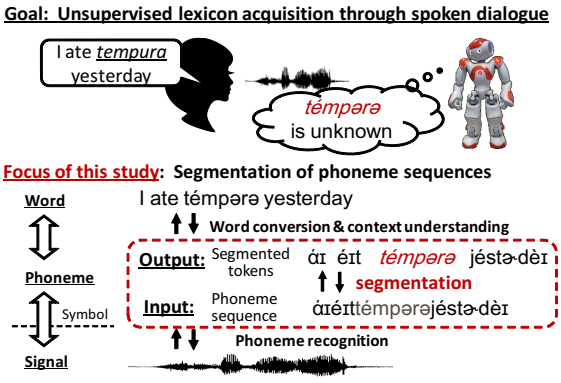

Figure 1: Our target problem

In this paper, we focus on the phoneme-level representation of utterance – not in the signal-level – to relax the problem, which results in an issue with phoneme sequence segmentation. The *segmentation* converts an input phoneme sequence into segmented sub-sequences corresponding to individual words. The focus of our study is illustrated in Figure 1. When the robot listens to a human utterance that includes an unknown word like "tempura," the robot will estimate the unknown segment "tempura" by the iterative search based on trial-and-error of many hypotheses among different layers. Here, the phoneme sequence of an unknown word is necessary as an intermediate representation between signals and lexicon because we cannot directly obtain the spelling of an unknown word just from sound information. Note that our assumption of a phoneme sequence given is partly supported by the high accuracy of state-of-the-art speech and phoneme recognition (Dahl et al., 2012; Hinton et al., 2012; Seide et al., 2011a,b).

Approaches based on Bayesian nonparametrics are promising methods to achieve lexical acquisition from unsegmented characters or phonemes. These methods estimate the *segmentation labels* of

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 243–252,
Taipei, Taiwan, November 27 – December 1, 2017

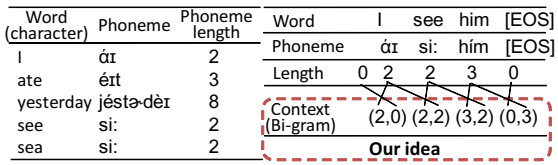

Figure 2: Example of phoneme-length and its context

phonemes corresponding to words with an unsupervised manner. The label represents the boundary of each *word*. Mochihashi *et al.* proposed the nested Pitman-Yor language model (NPYLM) (Mochihashi et al., 2009), or Pitman-Yor semi-Markov model (PYSMM) in other words. The model achieved high computational efficiency and high segmentation accuracy compared with a previous method based on the hierarchical Dirichlet process using simple Gibbs sampling (Goldwater et al., 2006). Uchiumi *et al.* also proposed a method that estimates the segmentation labels and part-of-speech tagging of words at the same time based on Pitman-Yor hidden semi-Markov models (PYHSMM) for character-level segmentation (Murphy, 2002; Uchiumi et al., 2015). PYHSMM has not been applied to the segmentation of phoneme sequences.

The difficulty with phoneme sequence segmentation is insufficient cues to distinguish or predict the context and segmentation labels. For example, the homophones are improperly treated as single entry and their different contexts are also confused in phoneme-level segmentation. This is a similar situation with the homographs in character-level segmentation, but it occurs much more frequently in phoneme-level segmentation, resulting in more serious problem. Although NPYLM and PYHSMM have been applied to character-level segmentation, they do not utilize cues useful for phoneme-level segmentation. We need to determine such useful cues to achieve accurate segmentation of phoneme sequences. Note that the performance comparison of NPYLM and PYHSMM methods in phoneme-level segmentation have not been conducted. We believe that comparing these methods on the basis of phoneme sequences is also useful for further improvement of the model.

We propose a *phoneme-length* context model for segmentation, which was not used in the NPYLM and PYHSMM. Note that the *length* is not the *duration* of a phoneme (the phoneme 'a' continues for three frames in the time axis, for example), which is used in signal-level segmentation (Lee and Glass, 2012). Figure 2 illustrates phoneme length and its contexts in the case of bi-grams. The phoneme sequence of 'see' and 'sea' is same and both lengths are two as shown in the left side of Fig. 2. The context of phoneme length is the sequence of these lengths. For example, if 'him' succeeds to 'see', the bi-gram of phoneme length is the pair of 3 and 2 where 3 is the length of phonemes 'him' and 2 is the length of phonemes 'si:'. We denote the pair as (3, 2) as shown in the right side of Fig. 2. Since the length of each segmented phoneme also depends on the previously segmented phonemes, this context represents one aspect of parts of speech. For example, the phoneme-length context captures the tendency that the length of the adposition is usually short and the length of the succeeding segment will be relatively long. We expect the phoneme-length context to be another cue for segmentation because the phoneme-length context is more abstract than word-level context. This phoneme-length model is expected to capture a rhythmic aspect of language.

We model the phoneme-length context as a prior probability distribution of sequential segmentation labels. This is because the probability distribution is expected to control how long each segmented phoneme becomes. Since the joint prior probability distribution of sequential segmentation labels were decomposed into factorized probabilities like N-gram, the phoneme-length model follows the Markov process and has a transition probability. The transition probability is also modeled and smoothed by using the Pitman-Yor N-gram model as other language models did. Our method, using NPYLM and PYHSMM, is evaluated by using a conversational corpus in English and Japanese in terms of the F-measures of the estimated segmentation labels. Because the corpus contains fillers and hesitations, the property of utterance used for evaluation matches our research purpose.

1.2 Related Work

There are several approaches based on Bayesian nonparametrics to achieve lexical acquisition from raw audio signals (Neubig et al., 2010; Lee and Glass, 2012; Kamper et al., 2016; Taniguchi et al., 2016), unsegmented phonemes, or words (Mochihashi et al., 2009;

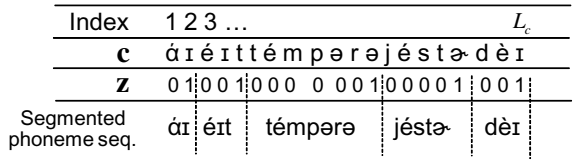

Figure 3: Word segmentation

Goldwater et al., 2009; Elsner et al., 2013; Uchiumi et al., 2015). The lexical acquisition technique is necessary in other areas, such as dialogue system that acquires knowledge through dialogue (Ono et al., 2016).

The advantages of Bayesian approach compared with other approaches (Kuo et al., 2007; Räsänen et al., 2015) are that a) the number of words in the system's vocabulary can be increased automatically in accordance with the amount of data and b) the semi-supervised learning of segmentation labels is easy to apply to utilize our knowledge of language. A typical estimation procedure is a Gibbs-sampling-based iteration of 1) the estimation of borders (segmentation labels) given a temporal N-gram language model (Goodman, 2001) and 2) the estimation of an N-gram language model given the temporal segmentation labels.

2 Unsupervised Segmentation and Baselines

We explain the overview and segmentation algorithm of NPYLM and PYHSMM as baseline methods. PYHSMM is an extended model of NPYLM to estimate the part-of-speech tagging of segmented words at the same time.

2.1 Overview

The unsupervised segmentation problem is defined finding the latent segmentation labels $\mathbf{z} = [z_1, ..., z_{L_c}]^T$ that correspond to each phoneme in the phoneme sequence $\mathbf{c} = [c_1, ..., c_{L_c}]^T$ with length L_c. If the binary label $z_i = 1$, the phoneme sequence is separated after the phoneme c_i. Figure 3 illustrates the role of $\mathbf{z}$. Other latent parameters $\mathbf{m} = [m_1, ..., m_{L_m}]$ with length L_m are also introduced to represent part of speech labels for each segmented phoneme sequences if necessary. The number of classes of part of speech label, M, is defined in advance of this study. The latent parameters are estimated by maximizing the following probabilities:

$$\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{c}) \propto p(\mathbf{c}|\mathbf{z}) \quad \text{or} \tag{1}$$

$$\arg\max_{\mathbf{z},\mathbf{m}} p(\mathbf{z}, \mathbf{m}|\mathbf{c}) \propto p(\mathbf{c}|\mathbf{z}, \mathbf{m})p(\mathbf{m}), \tag{2}$$

where NPYLM uses Eq. (1) and PYHSMM uses Eq. (2). The definition of each likelihood, such as $p(\mathbf{c}|\mathbf{z})$ and $p(\mathbf{c}|\mathbf{z}, \mathbf{m})$, is important. Because the border of phonemes $\mathbf{z}$ is given in these models, the likelihood can be factorized like N-gram probability. For example, the likelihood can be factorized as $p(c_{i+1}, ..., c_N|c_1, ..., c_i, \mathbf{z})p(c_1, ..., c_i|\mathbf{z})$, where the phoneme segments are considered to be two *word* segments $w_1 = c_1...c_i$ and $w_2 = c_{i+1}...c_{L_c}$. This N-gram modeling is also adopted to decompose the part-of-speech label $\mathbf{m}$, and this controls the grammar and the number of words.

The nested hierarchical Pitman-Yor language model (NPYLM) is used to represent the factorized N-gram probability (Mochihashi et al., 2009). Here, we represent the context of N-gram as $\vec{h}$ and the depth of the hierarchical context tree of $\vec{h}$ as $|\vec{h}|$. Given the seating arrangement of customers that are represented by hidden variables $\vec{s}$ in the hierarchical Chinese restaurant process (CRP), the conditional probability of a word segment w with the context $\vec{h}$ is defined as follows:

$$p(w|\vec{s}, \vec{h}) = \frac{c_{\vec{h}w} - d_{|\vec{h}|}t_{\vec{h}w}}{c_{\vec{h}*} + \theta_{|\vec{h}|}} + \frac{\theta_{\vec{h}} + d_{|\vec{h}|}t_{\vec{h}*}}{c_{\vec{h}*} + \theta_{|\vec{h}|}} p(w|\vec{s}, \vec{h}'), \tag{3}$$

where $c_{\vec{h}w}$ is the count of word w at context $\vec{h}$, and $c_{\vec{h}*} = \sum_w c_{\vec{h}w}$ is its sum. $\vec{h}'$ is the reduced context of $\vec{h}$, in which the relationship $|\vec{h}'| = |\vec{h}| - 1$ exists. $t_{\vec{h}w}$ is the number of tables at context $\vec{h}$, and $t_{\vec{h}*}$ is its sum. $\theta_{|\vec{h}|}$ and $d_{|\vec{h}|}$ are the common parameters of $\vec{h}$ with the same depth $|\vec{h}|$. Here, the uni-gram segment probability $p(w = c_i...c_j)$ is smoothed by the phoneme-level N-gram probability $p(c_i, ..., c_j) = p(c_j|c_i, ..., c_{j-1})p(c_i, ..., c_{j-1})$. Please see the work of Teh (2006) for the sampling algorithm of seating arrangement.

The segmentation labels $\mathbf{z}$ and other parameters such as the N-gram language model are updated iteratively. If the segmentation labels are given, we can calculate the statistics of the N-gram model. If the N-gram model is given, we can estimate the probability of the segmentation labels. Because the update of the N-gram model is well known, we explain the update of the estimation of the segmentation labels in the latter parts of this section.

Algorithm 1 Backward sampling: Θ represents parameter set

Require: $t \leftarrow N, i \leftarrow 0, w_0 \leftarrow \mathbb{E}$
  while $t > 0$ **do**
    Draw $k \propto p(w_i|c_{t-k+1}^t, \Theta)\alpha[t][k]$
    Set $w_i \leftarrow c_{t-k+1}^t$
    Set $t \leftarrow t-k, i \leftarrow i+1$
  end while

2.2 Inference for NPYLM

Mochihashi *et al.* (2009) proposed introducing the forward-backward inference to estimate the segmentation labels efficiently. This method uses a semi-Markov model, and it considers the problem as a sequential estimation of the hidden labels. The procedure consists of two steps: forward filtering and backward sampling.

The forward filtering calculates the forward probability that is used for the Bayesian learning of the hidden Markov model (HMM) (Scott, 2002). The following equation denotes $\alpha[t][k]$ as the probability of generating the partial phonemes $c_1, ..., c_t$ of $\mathbf{c}$ using the last k phonemes in the case of bi-grams.

$$\alpha[t][k] = \sum_{j=1}^{t-k} p(c_{t-k+1}^t|c_{t-k-j+1}^{t-k})\alpha[t-k][j], \quad (4)$$

where $\alpha[0][0] = 1$ and $c_n, ..., c_m = c_n^m$.

Backward sampling is achieved by drawing a phoneme segment w from the end of a sentence by using forward probability $\alpha[t][k]$. Because the end of sentence is represented by the special symbol $\mathbb{E}$, we can start sampling a word with the probability proportional to $p(\mathbb{E}|c_{N-k}^N)$. The algorithm is summarized in Alg. 1.

Note that we do not use the correction of phoneme-level N-gram probability based on the phoneme length using the Poisson distribution in the NPYLM. This is because the length property is embedded into our model naturally.

2.3 Inference for PYHSMM

The PYHSMM, which is an extended model of NPYLM, estimates the parts of speech of each segmented phoneme. We expect that the performance of PYHSMM is better than that of NPYLM. The forward probability, $\alpha[t][k][m]$, is newly introduced in the case of bi-grams, and the following equation denotes it as the probability of generating the partial phonemes $c_1, ..., c_t$ with part-of-speech m from the last k phonemes.

$$\alpha[t][k][m] = \sum_{j=1}^{t-k}\sum_{r=0}^{M} p(c_{t-k+1}^t|c_{t-k-j+1}^{t-k}, m) \\ p(m|r)\alpha[t-k][j][r] \quad (5)$$

where $p(m|r)$ is the transition probability of the latent parts of speech and assumed the hierarchical Pitman-Yor language model.

The algorithm of the backward sampling is similar to that of NPYLM. The parts of speech are sampled as well as the segmentation label. Because the end of the sentence and its parts of speech are represented using the special symbol $\mathbb{E}$, we can start sampling a word with the probability proportional to $p(\mathbb{E}|c_{N-k}^N, \mathbb{E})p(\mathbb{E}|m)$ like NPYLM. Note that *the computational cost of PYHSMM is larger than that of NPYLM* due to the search part-of-speech labels.

3 Analysis and Our Approach

We focus on the distribution of phoneme length to distinguish the confused contexts. If we have two different words with the same pronunciation, we can sometimes distinguish the phoneme representations of them on the basis of the lenght of the preceding or succeeding phoneme segments. The phoneme-length context will capture the tendency that the length of the adposition is usually short and the length of the succeeding segment will be relatively long.

Figure 4 illustrates the real phoneme-length distribution in the English and Japanese spoken-dialogue transcriptions used in our evaluation (Sec. 5.2). Given that the function $len(w)$ returns the phoneme length of word w, the matrices represent the bi-gram length probability $p(len(w_n)|len(w_{n-1}))$, and the horizontal axis is $len(w_{n-1})$ and the vertical axis is $len(w_n)$. w_n represents the n-th word in each sentences. The line graphs represent the uni-gram length probability $p(len(w_n))$. These probabilities were calculated on the basis of maximum likelihood estimation. *Verb* and *Noun* represent the phoneme-length probability of verbs and nouns in Japanese data, respectively. The definitions of bi-gram and uni-gram probability for *Verb* and *Noun* are $p(len(w_n)|len(w_{n-1}), pos(w_n))$ and $p(len(w_n|pos(w_n))$, where $pos(w)$ is a function that returns a part-of-speech tag of the word w.

We determined that the phoneme-length probability depends on 1) language, 2) context, and

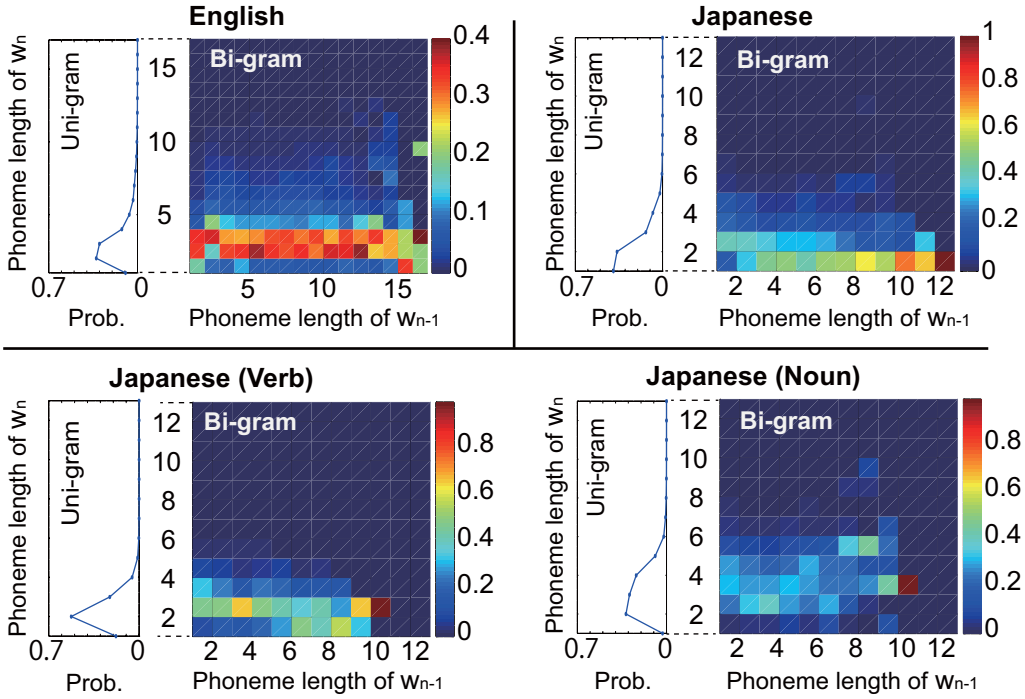

Figure 4: Real phoneme-length distribution

3) parts of speech. The bi-gram phoneme-length probabilities in English are relatively similar to each other but different from those in Japanese. Some bi-gram probabilities have several peaks, and they vary in accordance with parts of speech. If we utilize this information, we will achieve an accurate segmentation.

The straightforward approach to exploit phoneme-length information is to utilize the prior distribution of the segmentation labels $\mathbf{z}$ that is not used in either NPYLM and PYSHMM. Because the prior probability is considered to be the source that determines the length of each phoneme segment, embedding this prior into a model is expected to improve the segmentation performance. Therefore, we need to construct a model that considers the prior of the segmentation labels and should also reveal the performance of these models for phoneme-level word segmentation.

4 Phoneme-Length Context Model for Pitman-Yor Semi-Markov Models

We extend the NPYLM and PYHSMM to exploit the phoneme-length patterns of each phoneme segment. First, we explain our problem statement for unsupervised segmentation of phoneme sequences. Next, we derive the context model of the phoneme length and show the forward-backward algorithms for our extended NPYLM and PYHSMM.

4.1 Problem Statement and Model

We exploit the probability of phoneme length in estimating latent segmentation labels $\mathbf{z}$ and latent part-of-speech labels $\mathbf{m}$. The parameters are estimated by maximizing the following probabilities:

$$\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{c}) \propto p(\mathbf{c}|\mathbf{z})p(\mathbf{z}) \quad \text{or} \tag{6}$$

$$\arg\max_{\mathbf{z},\mathbf{m}} p(\mathbf{z},\mathbf{m}|\mathbf{c}) \propto p(\mathbf{c}|\mathbf{z},\mathbf{m})p(\mathbf{z}|\mathbf{m})p(\mathbf{m}). \tag{7}$$

The former objective function is for NPYLM, and the latter is for PYHSMM. The probabilities of segmentation labels $p(\mathbf{z})$ in Eq. (6) and $p(\mathbf{z}|\mathbf{m})$ in Eq. (7) are used in our objective functions. $p(\mathbf{z})$ is a prior probability distribution of segmentation labels $\mathbf{z}$ in Eq. (6).

We decompose each joint probability into N-gram probabilities. For an easy explanation of this decomposition, here we use the length of part-of-speech labels L_m and the correct segmentation labels as if these are given, which are actually searched for during training. The non-zero indices of segmentation labels $\mathbf{z}$ are represented by $\mathbf{g} = [g_1, ..., g_W]$, where W is the number of "true" phoneme segments. W equals L_m in the case of part-of-speech estimation. We also define $g'_i = g_i + 1$. The factorized models in the case

of bi-grams for NPYLM and PYHSMM are represented as follows:

$$p(\mathbf{c}|\mathbf{z}) = \prod_i p(c_{g'_{i-1}}^{g_i} | c_{g'_{i-2}}^{g_{i-1}}, z_{g'_{i-1}}^{g_i}, z_{g'_{i-2}}^{g_{i-1}}), \quad (8)$$

$$p(\mathbf{z}) = \prod_i p(z_{g'_{i-1}}^{g_i} | z_{g'_{i-2}}^{g_{i-1}}), \quad (9)$$

$$p(\mathbf{c}|\mathbf{z}, \mathbf{m}) = \prod_i p(c_{g'_{i-1}}^{g_i} | c_{g'_{i-2}}^{g_{i-1}}, z_{g'_{i-1}}^{g_i}, z_{g'_{i-2}}^{g_{i-1}}, m_i), \quad (10)$$

$$p(\mathbf{z}|\mathbf{m}) = \prod_i p(z_{g'_{i-1}}^{g_i} | z_{g'_{i-2}}^{g_{i-1}}, m_i), \quad (11)$$

$$p(\mathbf{m}) = \prod_i p(m_i | m_{i-1}), \quad (12)$$

where $p(m_i|m_{i-1})$ is a transition probability of latent part-of-speech labels, $z_{g'_{i-1}}^{g_i} = z_{g_{i-1}+1}, ..., z_{g_i}$ and $p(z_{g'_{i-1}}^{g_i} | z_{g'_{i-2}}^{g_{i-1}})$ and $p(z_{g'_{i-1}}^{g_i} | z_{g'_{i-2}}^{g_{i-1}}, m_i)$ are transition probabilities of the segmentation labels. The transition probability of segmentation labels is derived naturally. The latent variables for the seating arrangement of N-gram probability in Eq. (3) are omitted in these equations.

We design the transition probability of segmentation labels, such as $p(z_{g'_{i-1}}^{g_i} | z_{g'_{i-2}}^{g_{i-1}})$, to depend on the length of each phoneme segment. Because the length of each segment can be represented using the non-zero indices $\mathbf{g}$, the bi-gram transition probability is rewritten as

$$p(z_{g'_{i-1}}^{g_i} | z_{g'_{i-2}}^{g_{i-1}}) = p(g_i - g_{i-1} | g_{i-1} - g_{i-2}) \quad (13)$$

where m_{g_i} is omitted in the case of PYHSMM, and each integer, such as $g_i - g_{i-1}$, is considered as a symbol or label. This transition probability is the phoneme-length bi-gram probability as mentioned in Sec. 3, and it is also modeled by the hierarchical Pitman-Yor language model (HPYLM) not by a Poisson distribution (Mochihashi et al., 2009) because HPYLM is a count-based representation, which is appropriate for multimodal distribution. Such probability for duration modeling is also seen in (Kuo et al., 2007).

4.2 Inference

We derived the forward-backward algorithms to estimate the segmentation labels and part-of-speech labels. The inference for NPYLM is introduced first; then, the inference of PYHSMM is explained. Note that the segmentation label $\mathbf{z}$ and part-of-speech labels $\mathbf{m}$ are estimated simultaneously.

The forward probability $\alpha[t][k]$ of NPYLM with phoneme-length context is modified as follows.

$$\alpha[t][k] = \sum_{j=1}^{t-k} p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}, k, j) \, p(k|j) \alpha[t-k][j] \quad (14)$$

where $p(k|j)$ is a transition probability of the length of each phoneme segment. The forward probability is modified by the bi-gram probability of lengths. We can use $p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}, k)$ instead of $p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}, k, j)$ because information of length j is included in the phoneme sequence representation, $c_{t-k-j+1}^{t-k}$.

As with NPYLM, the context of phoneme lengths can be embedded into PYHSMM. The forward probability is also represented as

$$\alpha[t][k][m] = \sum_{j=1}^{t-k} \sum_{r=0}^{M} p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}, k, j, m) \, p(k|j, m) p(m|r) \alpha[t-k][j][r] \quad (15)$$

where m represents a part of speech. The forward probability is also biased by a transition probability of the length of each phoneme segment $p(k|j, m)$. We can also use $p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}, k, m)$. Because the number of parameters is large in this case of latent parts of speech, the convergence speed will degrade compared with NPYLM. Backward sampling of both cases is achieved in the same way as in NPYLM. The details are omitted due to space limitation.

4.3 Substitution

We substitute the conditional probability into simpler one by ignoring the dependency on length k in this work as follows:

$$p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}, k) := p(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}). \quad (16)$$

The probability should be ideally normalized only on the tokens that have length k, and this substitution makes a double count of the length information of token c_{t-k+1}^{t}. On the other hand, the difference between NPYLM and our proposed model in the inference is clear. The transition probability $p(k|j)$ is added in our model, and its implementation becomes simple. Our strict model will be evaluated in the future work.

We also use $p(c_{t-k+1}^{t}|c_{t-k-j+1}^{t-k}, m)$ instead of $p(c_{t-k+1}^{t}|c_{t-k-j+1}^{t-k}, k, m)$.

Note that there are several options on the back-off structure of the conditional probability $p(c_{t-k+1}^{t}|c_{t-k-j+1}^{t-k}, k)$. For example, the word uni-gram $p(c_{t-k+1}^{t}|k)$ might be smoothed by un-conditioned word uni-gram or by the k-conditioned character N-gram. If the amount of data is limited, the parameter estimation of the conditional character N-gram may fail. We adopt the existing model, NPYLM, as the structure in this work. The optimal structure should also be investigated in the future work.

5 Experiments

5.1 Evaluation Procedure

We evaluated each model by comparing the estimated segmented labels with the correct segmented original phoneme text (transcription of speech corpus). Utterances in the corpus were divided manually, and each utterance was treated as one sentence. The unsegmented phoneme text is generated by replacing the word-segmented transcription text into phoneme text with a dictionary and by removing whitespaces.

The criteria for the evaluations were the F-measures of the estimated lexicon set and segmentation label set. These F-measures are the harmonic mean of recall and precision; therefore, we could consider the recall and precision at the same time. The lexicon set after unsupervised segmentation was compared with that of the original segmented phoneme text. The estimated set of segmentation labels was also compared with that of the original phoneme text.

We used a development set to monitor the maximum performance of segmentation methods. The 1% data set was randomly selected from each test set, and its F-measure of segmentation labels was used to determine the epochs for calculating F-measures. Since we can obtain some correctly-transcribed text in a real situation, this evaluation process is reasonable. First, we ran each method over a sufficient number of epochs. Next, we calculated the F-measure of each development set's segmentation label and identified the 20-epoch section where the averaged F-measure was the highest. We calculated the F-measures of test sets that averaged over 20 epochs corresponding to the identified 20-epoch section. Note that each method is based on sampling and the maximum likelihood estimation sometimes does not match the segmentation on the basis of linguistic definitions.

Table 1: Parameters of experiment

	English	Japanese
Target text	SwitchBoard	CSJ
# of sentences	5,239	17,493
# of segments	88,127	132,900
# of phonemes	276,329	264,544
Vocab. size (word)	6,203	8,325
Vocab. size (phoneme)	5,422	6,589
Phoneme set	43	79

5.2 Data

We used two types of speech transcription in English and Japanese for evaluation. This is because the distribution of phoneme length also differs in languages as mentioned in Sec. 3.

We used the Switchboard-1 Telephone Speech Corpus (Godfrey et al., 1992) for the English set, which includes the transcription of conversational dialogue speech[1]. We selected 5,239 sentences from the session "ID 20," which included 88,127 word segments with 6,203 unique words. These words were converted into phonemes, totaling 276,329 phoneme characters. The vocabulary size in terms of phoneme representation was 5,422, and this was a unique number of phoneme sequences of words. For example, because the pronunciation of the words "see" and "sea" is the same "si:", the phoneme sequence "si:" is considered to be a unique vocabulary item. The phoneme set used in the English corpus included 43 phonemes in total including end-of-sentence symbols. The properties of the corpus are summarized in Table 1.

We used the Corpus of Spontaneous Japanese (CSJ) for the Japanese set, which is a collection of spoken dialogue recordings and their transcriptions (Maekawa, 2003). We used 17,493 sentences, including 132,900 word segments with 8,325 of them being unique words. The phoneme set for Japanese includes the combination of consonants and vowels and almost completely corresponds to *"katakana"* in Japanese to remove redundancy. The words were also transformed into phonemes ("*katakana*"), resulting in 264,544 of them. The vocabulary size in terms of phoneme representation was 6,589, and this was the unique number of phoneme sequences of words. The phoneme set used in the Japanese corpus included

[1] http://www.isip.msstate.edu/projects/switchboard/releases/

79 phonemes in total including end-of-sentence symbols. The properties of the corpus are also summarized in Table 1.

5.3 Parameter Settings

The parameters of NPYLM were the same for all models. The hyper parameters of the word language model were initialized as $\theta_{|h|} = 2.0$, $d_{|h|} = 0.5$, and the other parameters of prior probability distribution were all set to 1.0, such as the parameters of the beta distribution in NPYLM.

The hyper parameters of the phoneme (character) language, part-of-speech model and length model were the same as those in the language model. We set the maximum length of the phoneme sequence L_c to 10 due to the computational complexity. The number of classes of part of speech label, M, was set to 4 due to the small corpus size and computational cost. The initial labels of parts of speech were initialized randomly within the number of classes.

5.4 Results

The maximum F-measures of the lexicon and segmentation are listed in Tables 2 and 3 for the English and Japanese test sets. The notations *Lex.* and *Seg.* represent the F-measures of the lexicon and segmentation, respectively. *NPYLM-D* denotes the proposed NPYLM with our phoneme-length context model in Table 2, and *PYHSMM-D* denotes the proposed PYHSMM with our context model in Table 3.

The F-measures of the proposed NPYLM-D outperformed the NPYLM for both the English and Japanese test sets as shown in Table 2. The improvements in the Japanese corpus, 0.067 (Lex.) and 0.045 (Seg.), were larger than those in the English corpus, 0.003 (Lex.) and 0.01 (Seg.). This is because the bi-gram probability of phoneme length varies more in Japanese than in English, and the NPYLM-D could capture such tendencies. The NPYLM does not use any information other than the context of a segmented phoneme sequence. Therefore, the length model is useful to model the phoneme-level features. The lower performance of NPYLM-D after convergence might be caused by the conditional probability substitution and its double-count of length information.

The F-measures of the proposed PYHSMM-D were worse than those of PYHSMM for both the English and Japanese test sets as shown in Table 3. The performances of these methods were worse than those of NPYLM and NPYLM-D. The reasons for this are due to 1) the smaller amount of data to treat latent context variable $\mathbf{m}$ and 2) the overlap of *contextual information* between the phoneme length and the latent variable $\mathbf{m}$. Since the latent variable $\mathbf{m}$ represents the class of context, a sufficient amount of data will be required for achieving stable performance compared with NPYLM. Moreover, the context information represented by the latent variable $\mathbf{m}$ possibly includes our phoneme length context. Thus, it might be difficult for the PYHSMM-D to separate these two contexts in the case of completely unsupervised training. These problems may be solved in the *semi-supervised* case where we exploit pre-labeled data with already segmented words and their tagged parts of speech.

Table 2: F-measures of segmentation by NPYLM and NPYLM-D

		NPYLM (baseline)	NPYLM-D (proposed)
English	Lex.	0.602	**0.605**
	Seg.	0.897	**0.907**
Japanese	Lex.	0.344	**0.411**
	Seg.	0.748	**0.793**

Table 3: F-measures of segmentation by PYHSMM and PYHSMM-D

		PYHSMM (baseline)	PYHSMM-D (proposed)
English	Lex.	0.528	0.471
	Seg.	0.825	0.788
Japanese	Lex.	0.202	0.158
	Seg.	0.499	0.437

The F-measures of segmentation during training for the English and Japanese test sets are shown in Figures 5 and 6, respectively. The horizontal axis represents the epoch of Gibbs sampling, and the vertical axis represents the F-measures for the segmentation label set. The Gibbs sampling was stopped after at least ten days. Note that the F-measure of segmentation does not necessarily correlate with the likelihood and all methods were based on stochastic segmentation.

The F-measures of NPYLM and PYHSMM for the English and Japanese corpora improved in proportion to the number of epochs, and those of PYHSMM and PYHSMM-D did not converge as shown in Figures 5 and 6. Because the number of hidden parameters of PYHSMM and PYHSMM-D was large, their convergence speeds were slow.

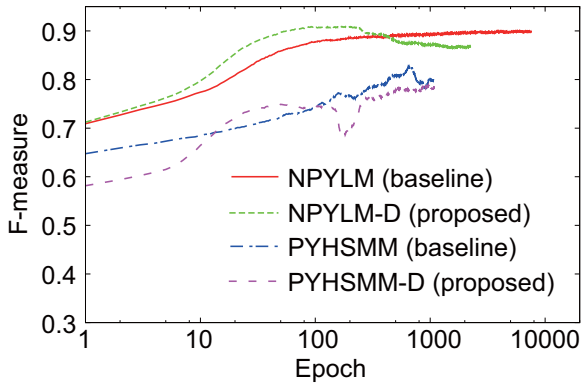

Figure 5: F-measures of segmentation for English test set

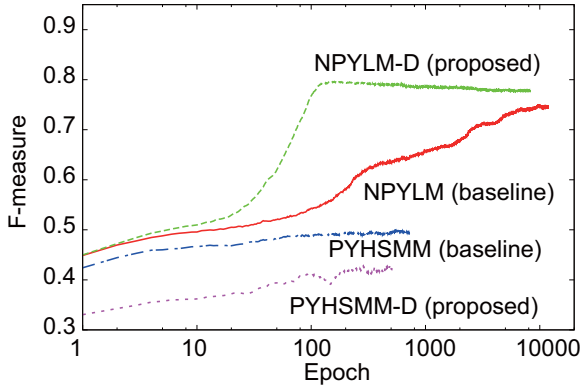

Figure 6: F-measures of segmentation for Japanese test set

The proposed NPYLM-D had peak F-measures at 100–200 epochs, and they went down and converged as the number of epochs increased. This was mainly because the phoneme-length context model accelerated the segmentation in the early epochs, and the small number of observations and un-supervised condition caused over-fitting. Therefore, the performance of the NPYLM-D is expected to be improved by using a larger corpus or by optimizing hyper parameters to match the actual prior probability of phoneme length.

The semi-supervised training would be effective for segmentation in practical use because it matches the actual use case. Evaluating the performance under such a condition is planned for future work. We expect the PYHSMM-D to work well more after 1000 epochs, but it requires a large computational cost. Its results could also be improved with a larger corpus and semi-supervised condition.

6 Conclusions

Unsupervised segmentation of phoneme sequences is an essential process to obtain unknown words during spoken dialogues with users. The PYSMM is a promising model to achieve unsupervised segmentation, but its performance degrades when it is applied to phoneme-level word segmentation. We proposed a phoneme-length context model for PYSMM to give a helpful cue at the phoneme-level and to predict succeeding phoneme segmentation more accurately. Our experiments showed that the peak performances with our context model outperformed those without such a context model by 0.045 at most in terms of F-measures of estimated segmentation labels.

There are the several future works on 1) optimization of parameters, 2) evaluation of semi-supervised training and other languages, and 3) improvement of our model and inference method. To further improve our method, we must investigate the hyper parameters setting for the estimation of segmentation labels to operate efficiently. The semi-supervised training will also improve the performance of our method. We will evaluate our method not only for English and Japanese but also other languages, such as African languages because the rhythmic information might be vivid. The performance of our strict model and the optimal back-off structure should be investigated to reveal the limitation of our model. The modification of inference algorithm will be required due to the computational efficiency when we use longer context information more than tri-gram. Since the rhythm information of latent segmentation might not be captured well by bi-grams, the challenge of using longer context is one of the important issues for our purpose.

Acknowledgments

We thank anonymous reviewers for their insightful comments. This work was supported partly by JSPS KAKENHI Grant Number JP16H02869.

References

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. 2013. A joint learning model of word

segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 42–54.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceesings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 673–680.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine*, 29(6):82–97.

Herman Kamper, Aren Jansen, and Sharon Goldwater. 2016. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 24(4):669–679.

Jen-Wei Kuo, Hung-Yi Lo, and Hsin-Min Wang. 2007. Improved HMM/SVM methods for automatic phoneme segmentation. In *Proceedings of Interspeech*, pages 2057–2060.

Chia-ying Lee and James Glass. 2012. A nonparametric bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 40–49. Association for Computational Linguistics.

Kikuo Maekawa. 2003. Corpus of spontaneous Japanese: Its design and evaluation. In *Inproceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108.

Kevin P Murphy. 2002. Hidden semi-Markov models.

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. 2010. Learning a language model from continuous speech. In *Proceesings of Interspeech*, pages 1053–1056.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2016. Toward lexical acquisition during dialogues through implicit confirmation for closed-domain chatbots. In *Second Workshop on Chatbots and Conversational Agent Technologies*.

Okko Räsänen, Gabriel Doyle, and Michael C Frank. 2015. Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Proceedings of Interspeech*, pages 3204–3208.

Steven L Scott. 2002. Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351.

Frank Seide, Gang Li, Xie Chen, and Dong Yu. 2011a. Feature engineering in context-dependent deep neural networks for conversational speech transaction. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 24–29.

Frank Seide, Gang Li, and Dong Yu. 2011b. Conversational speech transcription using context-dependent deep neural network. In *Proceedings of Interspeech*, pages 437–440.

Tadahiro Taniguchi, Shogo Nagasaka, and Ryo Nakashima. 2016. Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals. *IEEE Transactions on Cognitive and Developmental Systems*, 8(3):171–185.

Yee Whey Teh. 2006. A bayesian interpretation of interpolated kneser-ney. *Technical Report TRA2/06, School of Computing, NUS*.

Kei Uchiumi, Hiroshi Tsukahara, and Daichi Mochihashi. 2015. Inducing word and part-of-speech with Pitman-Yor hidden semi-Markov models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1774–1782.