Separating Precision and Mean in Dirichlet-Enhanced High-Order Markov Models

Rikiya Takahashi

IBM Tokyo Research Laboratory,
1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa 242-8502, Japan
rikiya@jp.ibm.com

Abstract. Robustly estimating the state-transition probabilities of high-order Markov processes is an essential task in many applications such as natural language modeling or protein sequence modeling. We propose a novel estimation algorithm called Hierarchical Separated Dirichlet Smoothing (HSDS), where Dirichlet distributions are hierarchically assumed to be the prior distributions of the state-transition probabilities. The key idea in HSDS is to *separate* the parameters of a Dirichlet distribution into the precision and mean, so that the precision depends on the context while the mean is given by the lower-order distribution. HSDS is designed to outperform Kneser-Ney smoothing especially when the number of states is small, where Kneser-Ney smoothing is currently known as the state-of-the-art technique for N-gram natural language models. Our experiments in protein sequence modeling showed the superiority of HSDS both in perplexity evaluation and classification tasks.

1 Introduction

To precisely predict or detect time-series sequences of discrete symbols, we desire robust inference techniques to estimate the state-transition probabilities in high-order Markov processes. Using state-transition probabilities for N-grams, a high-order Markov process is often used to model natural language [1], protein sequences [2], or the dynamics of consumers [3], where one state is assigned to each word, amino acid, or customer type, respectively. In these applications, the statistical robustness of the estimated state-transition probabilities often becomes a crucial issue for the predictive accuracy of the model, because the training data of the state-transition frequencies are limited. For example, word error rates in automatic speech recognition become high if we use N-gram language models trained with limited corpora.

In prior work, the state-of-the-art estimation techniques are not effective for cases when the number of states is small, such as a protein sequence, unless we use Markov Chain Monte Carlo (MCMC) methods. Generally, a standard strategy to robustly estimate the state-transition probabilities is to properly interpolate the probabilities of the N-grams, $(N-1)$-grams, and lower order distributions. In natural language modeling, the most advanced smoothing techniques currently used are Kneser-Ney smoothing [4] and its derivative versions [1]. The essence of Kneser-Ney smoothing and its derivatives is a modification of the state-transition

J.N. Kok et al. (Eds.): ECML 2007, LNAI 4701, pp. 382–393, 2007.

frequencies in calculating the lower order distributions, so that any frequency of a state-transition larger than one is reduced to one while the zero-frequency remains at zero. Such modifications of the frequencies are derived as a fast approximated inference of a hierarchical Poisson-Dirichlet (Pitman-Yor) process [5,6,7,8]. If we do not use that approximation, precisely estimating the parameters of hierarchical Pitman-Yor processes requires Gibbs sampling, which is a computationally intensive MCMC method. In addition, since the approximation is adequate only when the number of states is unbounded or sufficiently large, we are seeking another advanced estimation technique for when the number of states is bounded and small.

In this paper, we propose a novel technique to smooth the state-transition probabilities more effectively than Kneser-Ney smoothing when the number of states is small, and which does not require MCMC algorithms. We call our method Hierarchical Separated Dirichlet Smoothing (HSDS), because Dirichlet distributions are hierarchically assumed to be prior distributions of the state-transition probabilities. Our main idea is to *separate* the parameters of a Dirichlet distribution into a context-dependent precision and a mean given by the lower order distribution, and to estimate them alternately. Using the Dirichlet precision, we can quantify the effective frequency. Since the modified frequency adopted in Kneser-Ney smoothing is a special case of our effective frequency when the number of states is sufficiently large, HSDS can work flexibly when the number of states is small or large. In addition, since optimizing the parameters of a Dirichlet distribution does not require MCMC methods, HSDS runs relatively fast.

The rest of the paper is organized as follows. Section 2 introduces the hierarchical Dirichlet distributions that we use. Section 3 describes procedures to estimate the parameters of Dirichlet distributions and discusses when HSDS outperforms Kneser-Ney smoothing. Section 4 shows experimental results in the tasks of perplexity evaluation and classification, using natural language and protein sequence data. Section 5 concludes the paper.

2 Hierarchical Dirichlet Distributions for Prior

In this section, we introduce our custom Dirichlet distributions as the prior distributions of the state-transition probabilities, where our key idea is to impose different constraints on the precision and mean of the Dirichlet distribution. We hierarchically calculate the expectation of the state-transition probability on the posterior distribution, which is determined by the training data and the prior distribution. The mean of the Dirichlet distribution is given by lower-order distributions such as the $(N-1)$-gram models, which are more robust than the higher-order distributions. The precision of the Dirichlet distribution is a specific parameter for each context, to incorporate the numbers of unique states depending on that context. Our model is an extension of the hierarchical Dirichlet language model [9].

For a given discrete-state space $\mathcal{S}$ whose size is $|\mathcal{S}|$, assume we want to predict a prospective state s_N that will follow a state sequence $s_1, s_2, \cdots, s_{N-1}$ with

bounded length $N \geq 1$. Since s_N is a random variable, we need a model of $\Pr(s|h)$, the probability with which a state $s \in \mathcal{S}$ follows a $(N-1)$-length context $h \in \mathcal{S}^{N-1} \equiv \mathcal{S} \times \mathcal{S} \times \cdots \times \mathcal{S}$. We aim to estimate precise values of $p_{hs} \equiv \Pr(s|h)$ for each s and h from limited training data $\mathcal{D} = \left\{n_{hs}; s \in \mathcal{S}, h \in \mathcal{S}^{N-1}\right\}$, where n_{hs} is the frequency of state s that follows context h. A vector of estimated probabilities $\boldsymbol{p}_h$, whose i-th element p_{hi} is the probability with which the i-th state follows h, is defined as a random variable, because the estimated probability fluctuates around the true probability. For simplicity, hereinafter when a vector is defined with a bold face, such as $\boldsymbol{x}$, it is assumed that we simultaneously define its elements with a normal typeface of the same letter, such as x_s, where the element with the subscript s is a variable related to the state s.

Our aim is to estimate the expectation of $\boldsymbol{p}_h$ on the posterior distribution $P(\boldsymbol{p}_h|\mathcal{D})$. To compute a relevant posterior distribution, we need to specify a proper prior distribution $P(\boldsymbol{p}_h)$ for applying Bayes theorem. The expectation of the state-transition probability and the posterior distribution is given as

$$\langle p_{hs}|\mathcal{D}\rangle = \int_{\boldsymbol{p}_h} p_{hs} P(\boldsymbol{p}_h|\mathcal{D}) d\boldsymbol{p}_h \tag{1}$$

$$P(\boldsymbol{p}_h|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{p}_h)P(\boldsymbol{p}_h)}{\int_{\boldsymbol{p}_h} P(\mathcal{D}|\boldsymbol{p}_h)P(\boldsymbol{p}_h)d\boldsymbol{p}_h}, \tag{2}$$

where $\langle\cdot|\mathcal{D}\rangle$ is the expectation on the posterior distribution.

We assume a Dirichlet distribution in $P(\boldsymbol{p}_h)$ because its probability density function and its likelihood function are analytically tractable, and in order to avoid MCMC methods. This is contrast to adopting hierarchical Pitman-Yor processes that are more general stochastic processes but which require MCMC methods. With the parameters of the Dirichlet distribution $\boldsymbol{\phi}_h$ and the observed frequencies $\boldsymbol{n}_h$, the prior and posterior distributions are given as follows:

$$P(\boldsymbol{p}_h) = Dir\left(\boldsymbol{p}_h; \boldsymbol{\phi}_h\right) \equiv \frac{\Gamma\left(\sum_{s\in\mathcal{S}} \phi_{hs}\right)}{\prod_{s\in\mathcal{S}} \Gamma\left(\phi_{hs}\right)} \cdot \prod_{s\in\mathcal{S}} p_{hs}^{\phi_{hs}-1} \tag{3}$$

$$P(\boldsymbol{p}_h|\mathcal{D}) = Dir\left(\boldsymbol{p}_h; \boldsymbol{n}_h + \boldsymbol{\phi}_h\right) \equiv \frac{\Gamma\left(\sum_{s\in\mathcal{S}} n_{hs} + \phi_{hs}\right)}{\prod_{s\in\mathcal{S}} \Gamma\left(n_{hs} + \phi_{hs}\right)} \cdot \prod_{s\in\mathcal{S}} p_{hs}^{n_{hs}+\phi_{hs}-1} \tag{4}$$

Here we introduce the main idea of *separating* the parameters of the Dirichlet distribution into a precision and a mean that have different constraints from each other. We denote a truncated context, which is generated by removing the earliest state from h, by $\pi(h)$. We parameterize $\boldsymbol{\phi}_h$ as a product of the coefficient $\alpha_h = \sum_{s\in\mathcal{S}} \phi_{hs}$ and the normalized vector $\boldsymbol{\theta}_{\pi(h)}$. The expectation of the state transition probability p_{hs} on the posterior distribution is given as

$$\langle p_{hs}|\mathcal{D}\rangle = \frac{n_{hs} + \alpha_h \theta_{\pi(h)s}}{n_h + \alpha_h}. \tag{5}$$

Following Minka in [10], we call α_h the "Dirichlet precision" and $\boldsymbol{\theta}_{\pi(h)}$ the "Dirichlet mean".

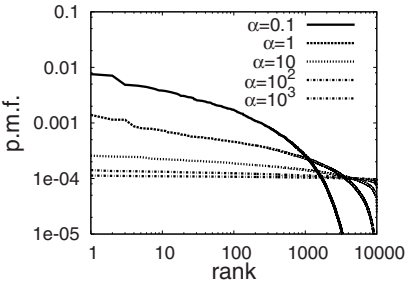


Fig. 1. Distributions of the states controlled by the Dirichlet precision α

The Dirichlet mean is hierarchically given by the expectation of the lower-order distribution, making use of the robustness of the lower-order distributions. We give $\boldsymbol{\theta}_{\pi(h)} = \langle \boldsymbol{p}_{\pi(h)} | \mathcal{D} \rangle$, assuming the prior $P(\boldsymbol{p}_{\pi(h)})$ by a Dirichlet distribution which has a precision $\alpha_{\pi(h)}$ and a mean $\boldsymbol{\theta}_{\pi(\pi(h))}$. Analogously, for each lower-order distribution from the $(N-2)$-gram to the 1-gram, the Dirichlet distribution is hierarchically assumed as its prior distribution. For the 1-gram, where h is empty, we define its Dirichlet mean by the 0-gram distribution $\boldsymbol{\theta}_0 = (1/|\mathcal{S}|, \cdots, 1/|\mathcal{S}|)^T$.

The Dirichlet precision depends on the context after which the different numbers of unique states will appear. Fig. 1 shows the multinomial distributions sorted in rank of probability with a log-log scale, where their parameters obey the 10,000-dimension symmetric Dirichlet distribution $Dir(\alpha/10000, \cdots, \alpha/10000)$. In Fig. 1, we see power-law distributions except when the states are extremely sporadic and that the higher Dirichlet precision α will yield larger numbers of unique states. Since a different number of unique states can follow from a different context h, we define the Dirichlet precision α_h for each context h. For example, in the 2-gram natural language model, α_h will be high if h is an article which does not strongly limit the following word, and α_h will be low if h is some specific verb such as "quantify", which tends to limit the following word.

HSDS can be regarded as an extension of the smoothing used in the hierarchical Dirichlet language model [9]. We believe MacKay and Peto were the first to use a Dirichlet distribution to smooth the probabilities of the 2-grams, where the prior distribution is given by $P(\boldsymbol{p}_h) = Dir(\boldsymbol{p}_h; \alpha \boldsymbol{\theta}_{\pi(h)h})$. Yet the original MacKay and Peto hierarchical Dirichlet language model was shown to be non-competitive with other smoothing techniques [8]. Since they discuss extensions to have the Dirichlet precision be context-dependent with classifying the contexts, HSDS is the first competitive method to extend the hierarchical Dirichlet language model.

3 Variational Inference by Effective Frequency

In this section, we present an algorithm to estimate the optimal Dirichlet precision and mean, and discuss when HSDS will outperform Kneser-Ney smoothing.

Our inference scheme is based on a variational approximation, where the infimum of the likelihood in a Dirichlet-multinomial distribution is maximized. The Dirichlet precision is optimized by a kind of Newton-Raphson method and the Dirichlet mean is calculated with the effective frequency, which is a frequency controlled by the Dirichlet precision. We explain when HSDS outperforms Kneser-Ney smoothing based on the meaning of the effective frequency.

First, we introduce the concept of the effective frequency by deriving the infimum of the likelihood of the training data. Since the observed frequencies $\boldsymbol{n}_h$ obey a Dirichlet-multinomial (Polya) distribution, Eq. (6) gives the likelihood of $\mathcal{D}$ under the set of hyperparameters $\Phi = \left\{\alpha_h, \boldsymbol{\theta}_{\pi(h)}; h \in \mathcal{S}^{N-1}\right\}$. We referred to [10] in deriving Inequality (7).

$$P(\mathcal{D}|\Phi) \propto \prod_h \frac{\Gamma(\alpha_h)}{\Gamma(n_h + \alpha_h)} \prod_{s:n_{hs}>0} \frac{\Gamma\left(n_{hs} + \alpha_h \theta_{\pi(h)s}\right)}{\Gamma\left(\alpha_h \theta_{\pi(h)s}\right)} \tag{6}$$

$$\geq \prod_h \frac{\Gamma(\bar{\alpha}_h)}{\Gamma(n_h + \bar{\alpha}_h)} \exp\left[\left(\Psi(n_h + \bar{\alpha}_h) - \Psi(\bar{\alpha}_h)\right)(\bar{\alpha}_h - \alpha_h)\right]$$

$$\prod_{s:n_{hs}>0} \left[\frac{\Gamma\left(n_{hs} + \bar{\alpha}_h \bar{\theta}_{\pi(h)s}\right)}{\Gamma\left(\bar{\alpha}_h \bar{\theta}_{\pi(h)s}\right)} \left(\bar{\alpha}_h \bar{\theta}_{\pi(h)s}\right)^{-\widetilde{n}_{hs}}\right] \left(\alpha_h \theta_{\pi(h)s}\right)^{\widetilde{n}_{hs}}, \tag{7}$$

where $\Psi(\cdot)$ denotes a *digamma* function such that $\Psi(x) \equiv \frac{\partial}{\partial x} \log \Gamma(x)$, and

$$\widetilde{n}_{hs} = \bar{\alpha}_h \bar{\theta}_{\pi(h)s} \left(\Psi\left(n_{hs} + \bar{\alpha}_h \bar{\theta}_{\pi(h)s}\right) - \Psi\left(\bar{\alpha}_h \bar{\theta}_{\pi(h)s}\right)\right). \tag{8}$$

We call $\widetilde{n}_{hs}$ the "effective frequency", because the infimum of the likelihood has the same formulation as a multinomial distribution for context h where the observed frequency of state s is $\widetilde{n}_{hs}$. Minka also discusses the effective frequency in [10], by differentiating the likelihood of the Polya distribution with respect to the Dirichlet mean. We explain the meaning of the effective frequency in Section 3.2. For convenience, we also define $\widetilde{n}_h \equiv \sum_{s \in \mathcal{S}} \widetilde{n}_{hs}$, $\widetilde{n}_{\pi(h)s} \equiv \sum_u \widetilde{n}_{hs}$ where u is the earliest state in context h and thus $h \equiv u\pi(h)$, and $\widetilde{n}_{\pi(h)} \equiv \sum_{s \in \mathcal{S}} \widetilde{n}_{\pi(h)s}$.

3.1 Estimating the Dirichlet Precision

Next, we estimate the Dirichlet precision as the expectation of α_h on an approximated posterior distribution. The procedure for estimation is divided into the following two cases. First, when $n_{hs} = 1$ for all s such that $n_{hs} > 0$, we initially set $\alpha_h = \infty$, which is equivalent to using only the state-transition probability of the $(N-1)$-gram[1]. Second, in all other cases, α_h is given by the expectation of a gamma distribution that approximates the posterior distribution of α_h. We assume the prior for α_h is a non-informative uniform distribution. Since the part of the likelihood related to α_h can be expressed as

$$P(\mathcal{D}|\alpha_h) \propto \exp\left[-\left(\Psi(n_h + \bar{\alpha}_h) - \Psi(\bar{\alpha}_h)\right)\alpha_h\right] \alpha_h^{\widetilde{n}_h}, \tag{9}$$

[1] If $\forall s, n_{hs} = 1$, then the exact likelihood expressed by Eq. (6) becomes a function of the Dirichlet mean alone. Therefore, the posterior distribution of α_h becomes the non-informative uniform distribution $U[0, \infty]$, whose expectation is infinity.

we can derive the approximated posterior $Q(\alpha_h|\mathcal{D})$ as

$$\begin{aligned} Q(\alpha_h|\mathcal{D}) &\propto P(\mathcal{D}|\alpha_h)P(\alpha_h) \\ &\propto Ga\left(\alpha_h; \widetilde{n}_h + 1, \Psi\left(n_h + \bar{\alpha}_h\right) - \Psi\left(\bar{\alpha}_h\right)\right), \end{aligned} \tag{10}$$

where we denote a gamma distribution by $Ga(\cdot,\cdot)$. The expectation of α_h on the approximated posterior $Q(\alpha_h|\mathcal{D})$ is given as

$$\langle \alpha_h|\mathcal{D}\rangle = \frac{\widetilde{n}_h + 1}{\Psi\left(n_h + \bar{\alpha}_h\right) - \Psi\left(\bar{\alpha}_h\right)}. \tag{11}$$

To calculate the optimal Dirichlet precision α_h^*, we assume $\langle \alpha_h|\mathcal{D}\rangle = \bar{\alpha}_h$, which means that the likelihood of the Polya distribution is approximated by the gamma distribution that has the same expectation. We can immediately derive the following equation, which we can solve quickly with the modified Newton-Raphson method proposed in [11].

$$\Psi\left(n_h + \alpha_h^*\right) - \Psi\left(\alpha_h^*\right) = \frac{1}{\alpha_h^*} + \sum_{s:n_{hs}>0} \theta_{\pi(h)s}\left[\Psi\left(n_{hs} + \alpha_h^*\theta_{\pi(h)s}\right) - \Psi\left(\alpha_h^*\theta_{\pi(h)s}\right)\right] \tag{12}$$

Note that the estimated α_h^* values tend to be underestimated when n_h is small, because the true posterior distribution of α_h has a heavier-tail than the gamma distribution. Based on several earlier experiments, we decided to multiply α_h^* by 2 if $\alpha_h^* > 10$. This is a simple heuristic rule, but it works for many datasets. A better estimation technique should be developed.

3.2 Estimating the Dirichlet Mean

The optimal Dirichlet mean $\theta^*_{\pi(h)s}$ is calculated from the effective frequency and the Dirichlet precision of the lower-order distributions. The terms related to $\theta_{\pi(h)s}$ in the infimum of the represented likelihood are also given by multinomial distributions that have effective frequencies as

$$Q(\mathcal{D}|\Phi) \propto \prod_{\pi(h)} \prod_{s:n_{hs}>0} \theta_{\pi(h)s}^{\widetilde{n}_{\pi(h)s}}. \tag{13}$$

Since we also assume a Dirichlet distribution in $P(\boldsymbol{p}_{\pi(h)})$, the optimal Dirichlet mean $\theta^*_{\pi(h)s}$ is given as

$$\theta^*_{\pi(h)s} = \frac{\widetilde{n}_{\pi(h)s} + \alpha_{\pi(h)}\theta_{\pi(\pi(h))s}}{\widetilde{n}_{\pi(h)} + \alpha_{\pi(h)}}. \tag{14}$$

Because the optimal Dirichlet precision and Dirichlet mean must be estimated iteratively, we summarized the computational procedure in Algorithm 1, except for the last heuristic multiplication for the Dirichlet precision.

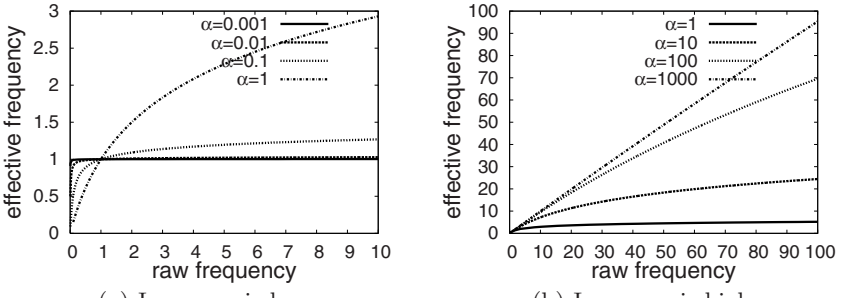
Algorithm 1. Estimating the Dirichlet precision and Dirichlet mean

```
Initialize all the parameters {α_h, θ_π(h)}.
repeat
  for n = N downto 1 do
    for all h ∈ S^{n-1} do
      if ∃s, n_hs > 1 then
        α_h ⇐ α*_h by solving Eq. (12).
      end if
      for all s ∈ S do
        Calculate ñ_hs by Eq. (8).
      end for
    end for
    if n ≥ 2 then
      for all h ∈ S^{n-1} do
        Update θ_π(h) by Eq. (14).
      end for
    end if
  end for
until all the parameters have converged.
```

3.3 Effects of the Effective Frequency

Finally, we discuss the cases when HSDS outperforms Kneser-Ney smoothing, by clarifying the meaning of the effective frequency. Fig. 2 shows the relationships between the effective frequency $\widetilde{n}$ and the raw frequency n, as functions of the Dirichlet precision α where $\widetilde{n} = \alpha\,(\Psi(\alpha + n) - \Psi(\alpha))$. When $\alpha \to 0$, the effective frequency converges to an indicator function of the raw frequency: $\widetilde{n} = 1$ if $n > 0$ and $\widetilde{n} = 0$ if $n = 0$ and it is the same as the modified frequency adopted in Kneser-Ney smoothing.

Fig. 2 and the actual effective frequency defined by Eq. (8) suggest that approximating the effective frequency by the modified frequency of Kneser-Ney smoothing is adequate only for s and h such that $\alpha_h \theta_{\pi(h)s}$ is low. When the



(a) In case α is low (b) In case α is high

Fig. 2. Effective frequencies with various values of the Dirichlet precision α

number of states is large, which is true for natural language modeling, the approximation is adequate because most of the values of $\theta_{\pi(h)s}$ are very low.

HSDS will outperform Kneser-Ney smoothing in two cases: when the number of states is small, or when the order of the Markov processes is high, in that $\alpha_h \theta_{\pi(h)s}$ is not too low in either case. First, if the number of states is small, some of the $\left\{\theta_{\pi(h)s}\right\}$ are not low. Second, if N is high and n_h is too small, α_h becomes a high value in estimating the Dirichlet precision. At such times, the effective frequency approaches the raw frequency. Intuitively, if the observed frequency of N-grams is too low, we should ignore the frequency of that N-grams and should just use the raw frequency of the $(N-1)$-grams.

4 Experiments

In this section, we experimentally show that HSDS outperforms Kneser-Ney smoothing when the number of states is small, by comparing the results for two different types of datasets: a natural language corpus and some protein sequence data. A natural language is chosen as a sequence with large number of states, because natural languages have large and potentially infinite vocabularies. Protein sequences are chosen as sequences with small number of states, because any protein sequence consists of only 20 types of amino acids, which means the number of states in a protein sequence N-gram is also limited to 20.

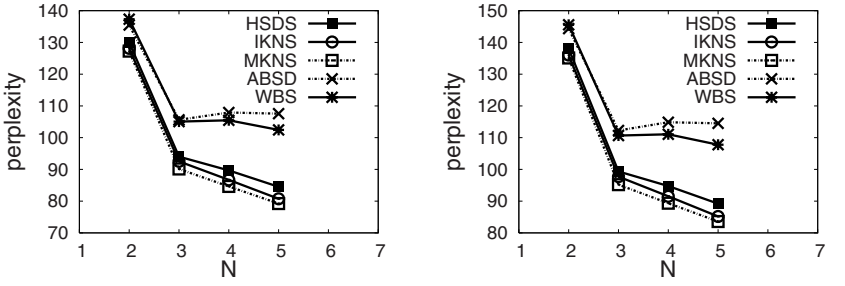
To evaluate the performance of each model, we focused on calculating the test-set perplexity, where its low values usually mean better predictive accuracies. In addition, we checked the results of classification tests for protein sequence modeling. For a K-length sequence $s_1^K \equiv s_1 s_2 \cdots s_K$, its perplexity evaluated by the N-gram model $\Theta = \left\{Pr(s|h), s \in \mathcal{S}, h \in \mathcal{S}^0, \mathcal{S}^1 \cup \cdots \cup \mathcal{S}^{N-1}\right\}$ is given as

$$PP\left(s_1^K|\Theta\right) = \exp\left[-\frac{1}{K}\sum_{k=1}^{K}\log Pr\left(s_k|s_{\max\{1,k-N+1\}},\cdots,s_{k-1}\right)\right]. \quad (15)$$

The other experimental conditions, which are common in natural language modeling and protein sequence modeling, are given below. After studying the numbers of unique N-grams in the training data, we decided to train the 2-, 3-, 4-, and 5-gram models. The 6-gram models were also trained for the protein sequence data. To compare the smoothing methods, we tested Hierarchical Separated Dirichlet Smoothing (HSDS), Interpolated Kneser-Ney Smoothing (IKNS), Modified Kneser-Ney Smoothing (MKNS), Absolute Discounting (ABSD) [12], and Witten-Bell smoothing (WBS) [13]. The smoothing methods except for IKNS and MKNS were selected to compare the performances broadly. The formulas used in IKNS and MKNS are described in [1], where we adopted the versions without cross-validation in estimating the discounting factors.

4.1 Natural Language Modeling

As natural language data, we used the `Reuters-21578` text categorization test collection [14], which is a popular English corpus mainly used for text



(a) When OOV words were included. (b) When OOV words were excluded.

Fig. 3. Test-set perplexity in `Reuters-21578` dataset

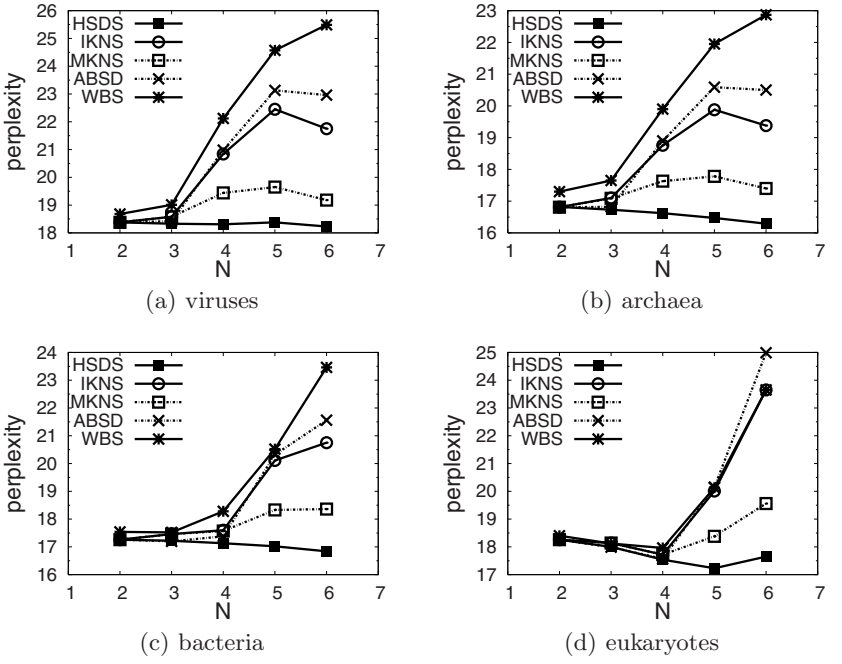
categorization research. By extracting all of the text, we prepared 172,900 sentences that consisted of a total of 2,804,960 words, and divided them into 162,900 training sentences and 10,000 test sentences. The training data had 118,602 unique words that appeared at least once, and we chose the most frequent 20,000 words as the vocabulary set. We calculated test-set perplexity both when out-of-vocabulary (OOV) words were included and excluded. When the OOV words were included, we replaced all of the OOV words with the same special token.

Fig. 3 shows each model's test-set perplexity and HSDS is inferior to both IKNS and MKNS. We think that the relatively weak performance of HSDS is because the Dirichlet distribution cannot precisely capture the power-law in the frequencies of the words. As shown in Fig. 1, the Dirichlet distribution cannot represent the heavy-tail of the frequencies of the words, while the Pitman-Yor process and Kneser-Ney smoothing can control the exponent of the power-law [8], which is important for the distribution within a potentially infinite vocabulary.

Still, HSDS outperformed the other smoothing techniques except for IKNS and MKNS. We think that the effective frequency in HSDS worked more effectively than the raw-frequencies, in calculating the lower-order distributions.

4.2 Protein Sequence Modeling

For protein sequence data, we performed classification tests as well as a perplexity evaluation, using the `DBsubloc` dataset [15]. Though `DBsubloc` is a protein database mainly used for protein subcellular localization, because of the amount of available data, we only classified the unlabeled data into one of 4 types of organisms: viruses, archaea, bacteria, and eukaryotes. After dividing the non-redundant dataset into training data and test data, we independently trained 4 types of N-gram models. In the training data, the numbers of unique sequences were 1,082 for the viruses, 1,131 for the archaea, 9,701 for the bacteria, and 18,043 for the eukaryotes. The test data consisted of 100 unique sequences for each organism, where their numbers of amino acids were 43,990 for the viruses, 26,455 for the archaea, 29,075 for the bacteria, and 62,286 for the eukaryotes.



(a) viruses (b) archaea

(c) bacteria (d) eukaryotes

Fig. 4. Test-set perplexity in `DBsubloc` dataset

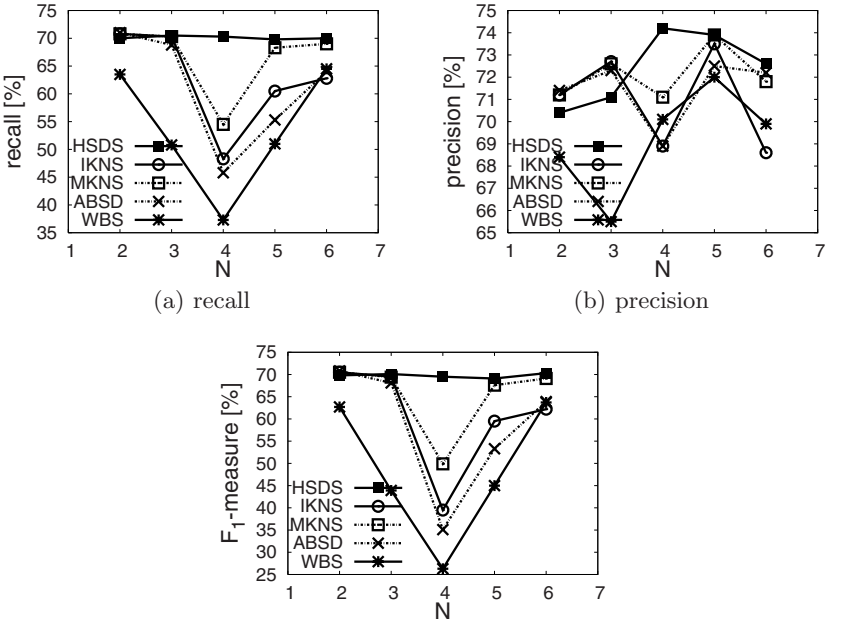
Perplexity Evaluation. In the perplexity evaluation task, each test-set was evaluated by a model of the same organism. i.e. The viruses test data was evaluated with viruses model. Fig. 4 shows the test-set perplexity for each organism.

HSDS achieved the lowest test-set perplexity, and its performance was slightly improved even when N became larger, while the other smoothing techniques had worse performances. As mentioned in Section 3.3, in protein sequence modeling, the effective frequency seemed to work more effectively than the modified frequency adopted in Kneser-Ney smoothing.

Classification. In the classification task, we made unlabeled data by removing the labels from all of the test data, and classified the unlabeled data into one of the 4 organism types using a naive Bayes classifier. Let c_i be one of the 4 organism types. For a sequence s_1^K, the organism type of the sequence $c(s_1^K)$ was determined as

$$c(s_1^K) = \arg\max_{c_i} P(s_1^N|c_i)P(c_i), \tag{16}$$

where $P(s_1^N|c_i)$ was calculated by the trained N-gram model of the organism type c_i, and $\forall c_i, P(c_i) = 0.25$. For each organism, we calculated the recall, precision, and F_1-measure. We used the arithmetic average of the 4 organism types as a performance metric of our multi-class classification.



(a) recall (b) precision

(c) F_1-measure

Fig. 5. Performances in classifying 4 organism-types

The results also show that HSDS is stable as N increases. Fig. 5 shows the averages of the recall, precision, and F_1-measure for the 4 organism types when the order of the N-gram changes. As in the perplexity evaluation, the performance of HSDS was stable even as N increased, while the performances of the other methods peaked for the 2-gram models. If we only look at the absolute performance, the 2-gram model with ABSD recorded the highest F_1-measure, but the other methods also recorded almost the same performances in 2-gram models.

5 Conclusion

We proposed a smoothing method for probabilistic N-gram models, which we named Hierarchical Separated Dirichlet Smoothing (HSDS). We hierarchically assumed a Dirichlet distribution to be a prior distribution of the state-transition probabilities, and separated the parameters of a Dirichlet distribution into precision and mean. The context-specific Dirichlet precision can reflect the context-dependent number of unique states, and the Dirichlet mean based on the effective frequencies gives appropriate lower-order distributions. Theoretically and experimentally, HSDS was shown to outperform Kneser-Ney smoothing when the number of states is small and N increases.

In the future, we will extend our context-specific formulation for more general stochastic process models such as the hierarchical Pitman-Yor processes, to more precisely incorporate the effects of the power-law in the observed frequencies.

Acknowledgment

The author wishes to thank Gakuto Kurata and Hisashi Kashima for many fruitful discussions about Kneser-Ney smoothing and other related topics.

References

1. Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard Computer Science (1998)
2. Ganapathiraju, M., Manoharan, V., Klein-Seetharaman, J.: BLMT: Statistical sequence analysis using n-grams. Applied Bioinformatics 3 (November 2004)
3. Netzer, O., Lattin, J.M., Srinivasan, V.: A Hidden Markov Model of Customer Relationship Dynamics. Stanford GSB Research Paper (July 2005)
4. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 181–184 (May 1995)
5. Pitman, J., Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. The Annals of Probability 25(2), 855–900 (1997)
6. Goldwater, S., Griffiths, T., Johnson, M.: Interpolating between types and tokens by estimating power-law generators. In: Advances in Neural Information Processing Systems (NIPS), vol. 18 (2006)
7. Teh, Y.W.: A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore (2006)
8. Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 44 (2006)
9. MacKay, D.J.C., Peto, L.: A hierarchical Dirichlet language model. Natural Language Engineering 1(3), 1–19 (1994)
10. Minka, T.: Estimating a Dirichlet distribution. Technical report, Microsoft Research (2003)
11. Minka, T.: Beyond Newton's method. Technical report, Microsoft Research (2000)
12. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependences in stochastic language modeling. Computer, Speech, and Language 8, 1–38 (1994)
13. Witten, I.H., Bell, T.C.: The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Transactions on Information Theory 37(4), 1085–1094 (1991)
14. Lewis, D.D.: Reuters-21578 text categorization test collection distribution 1.0 (1997) Available at `http://www.daviddlewis.com/resources/testcollections/reuters21578/`
15. Guo, T., Sun, Z.: Dbsubloc: Database of protein subcellular localization (2005) Available at `http://www.bioinfo.tsinghua.edu.cn/~guotao/`