Hierarchical Pitman-Yor and Dirichlet Process for Language Model

Jen-Tzung Chien and Ying-Lan Chang

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 30010, ROC
jtchien@nctu.edu.tw

Abstract
This paper presents a nonparametric interpretation for modern language model based on the hierarchical Pitman-Yor and Dirichlet (HPYD) process. We propose the HPYD language model (HPYD-LM) which flexibly conducts backoff smoothing and topic clustering through Bayesian nonparametric learning. The nonparametric priors of backoff n-grams and latent topics are tightly coupled in a compound process. A hybrid probability measure is drawn to build the smoothed topic-based LM. The model structure is automatically determined from training data. A new Chinese restaurant scenario is proposed to implement HPYD-LM via Gibbs sampling. This process reflects the power-law property and extracts the semantic topics from natural language. The superiority of HPYD-LM to the related LMs is demonstrated by the experiments on different corpora in terms of perplexity and word error rate.

Index Terms: language model, backoff model, topic model, Bayesian learning

1. Introduction
Statistical language model (LM) plays an important role in many information systems including machine translation, document classification, writing correction, bio-informatics, and speech recognition. The LM \( p(W) \) based on n-gram aims to calculate the probability of a word string \( W \) by multiplying the probabilities of a predicted word \( w \) conditional on its preceding \( n - 1 \) words. In general, n-gram model suffers from the inadequacies of training data and long-distance information [7][16]. The modified Kneser-Ney (MKN) LM [6][12] was proposed to tackle the inadequate training data by recursively performing backoff scheme and interpolating with \((n-1)\)-grams. The backoff could be also conducted through a structural Bayesian modeling [25]. To compensate insufficient long-distance information, the topic-based language model [8][9][20][23] was constructed by combining large-span latent topic information [1]. An unsupervised LM adaptation was proposed to incorporate topic mixtures based on latent Dirichlet allocation (LDA) [2].

More recently, Bayesian nonparametric (BNP) learning [3] has been extensively studied in machine learning community. BNP methods flexibly infer the model complexity from data without assuming parametric prior and posterior distributions. Teh [22] proposed a BNP approach to backoff LM according to a hierarchical Pitman-Yor (PY) process [15]. Hierarchical PY (HPY) process draws the power-law distributions which is a striking property of natural languages [10]. HPY-LM was interpreted as the Bayesian extension of MKN-LM [11][22]. In [19], the class-based HPY-LM was established to characterize many-to-many mapping between words and classes for conversational speech. In [24], a doubly HPY-LM was proposed for LM adaptation. A shared LM was adapted to each domain which was represented by an individual HPY-LM. In [14], a nested HPY-LM was combined with dynamic programming for word segmentation. However, these HPY-LMs [11][14][22][24] did not explore topic information. In [17], a PY topic model was constructed but only for document modeling. In [13], HPY process was combined with a topic model for phrase modeling where the parametric LDA model was considered.

This paper presents the BNP learning for LM which allows model growing structurally as more data are observed. We propose a topic-based LM [5] according to the HPY process compound hierarchical Dirichlet process (HDP) [21]. Using this HPYD-LM, the integrated nonparametric priors are constructed to draw topic-dependent backoff n-grams and simultaneously combine them into a mixture model of topical n-grams. HDP and HPY are tightly integrated to draw HPYD-LM with power-law property and coherent topic information.

2. Prior Works

2.1. Topic-based language model

Topic-based LM [9] was proposed to capture long-range word dependencies through discovery of latent topics. The resulting n-gram is expressed by

\[
p(w_i|w_{i-n+1}^{i-1}) = \sum_{z_i=k} p(w_i|w_{i-n+1}^{i-1},z_i)p(z_i|w_{i-n+1}^{i-1})
\]

where \( z_i = k \) denotes the topic label of word \( w_i \) from \( K \) topics and \( p(z_i|w_{i-n+1}^{i-1}) \) denotes the topic proportion given history words \( h = w_{i-n+1}^{i-1} \). The topic-based unigrams and bigrams are calculated by \( p(w_i) = \sum_{z_i} p(w_i|z_i)p(z_i) \) and \( p(w_i|w_{i-1}) = \sum_{z_i} p(w_i|w_{i-1},z_i)p(z_i|w_{i-1}) \), respectively. The maximum likelihood estimates of topic-based LM are calculated according to the expectation-maximization (EM) algorithm. In [8], a cache Dirichlet class LM (cDC-LM) was estimated by a variational Bayes EM procedure where the lower bound of log marginal likelihood was maximized. The likelihood was marginalized over latent classes or topics which were represented by Dirichlet distributions. Different from class-based LM [4] based on hard-clustering, the topic-based LMs [8][9] perform soft-clustering over all topics. Nevertheless, these methods calculated the parametric mixture models where the number of topics \( K \) was fixed. BNP learning aims to relax this assumption and conduct the structural learning.

2.2. Bayesian nonparametric learning

BNP learning using HPY [15][22] and HDP [21] have been proposed to infer LM and document model, respectively. HPY process [22] was developed to draw nonparametric n-gram
Each document or group of documents is treated as a single compound process. Starting from the uniform seed measure \( H_0 \), we draw a global topic measure \( G_0 \) from a global DP \( \pi_{G0} \) which represents a single set of mixtures is shared across documents. Then, each document or group \( d \) is drawn from a DP \( G_d \) which determines how much a mixture component from a shared mixture model contributes to that document. The global base measure of \( G_0 \), denoted by \( \theta_0 \), is self-sampled from a global DP \( G_0 \) which ensures that a single set of mixtures is shared across documents. The strength parameter \( \theta_0 \) determines the proportion of a mixture in a document \( d \). The document distribution \( G_d \) over \( d \) is generated by

\[
G_d \sim \text{DP}(\gamma_0, G_0), \quad G_0 \sim \text{DP}(\theta, G_0)
\]

where \( \gamma_0 \) and \( G_0 \) denote the strength parameter and base measure of \( G_0 \), respectively. HDP is developed to represent “a bag of words” from a set of documents through nonparametric prior \( G_0 \). The sequence of words is not characterized by HDP.

### 3. HPYD Language Model

This paper presents a hierarchical Pitman-Yor and Dirichlet language model (HPYD-LM) which jointly conducts backoff smoothing and topic modeling through BNP learning. The number of topics \( K \) is learnt from data. This model is different from Bayesian class-based LM \( \text{p}(w|h,c)p(c|h) \) [19] where two HPY processes were developed to separately sample mixture probabilities \( p(w|h,c) \) and \( p(c|h) \) while number of classes \( C \) was fixed. In what follows, HDPY process is shown as a single compound process.

#### 3.1. HPYD process

HPYD-LM assumes that \( n \)-gram is expressed by a nonparametric topic mixture model. HPYD process is described as follows. Starting from the uniform seed measure \( H_0 \), we draw a global topic measure \( G_0 \) from a DP \( \text{DP}(\gamma_0, H_0) \). The topic-dependent unigram \( H_{d,w} \) with topic assignment \( w \) is sampled by \( H_{d,w} \sim \text{PY}(\theta_1, H_0) \) where \( G_0 \) is acted as a prior base measure. Next, \( H_{d,w} \) serves as a base measure for a DP to draw unigram probability \( G_{w_{i}} \sim \text{DP}(\gamma_1, H_{d,w}) \). Using \( G_{w_{i}} \) as a prior measure, we draw topic-dependent bigram by using \( \text{PY} \) process \( H_{d,w_{i-1}w_{i}} \sim \text{PY}(\theta_2, H_{d,w_{i-1}}) \). This measure is again acted as a prior basis for a DP to draw bigram \( G_{w_{i-1}w_{i}} \sim \text{DP}(\gamma_2, H_{d,w_{i-1}}) \). Using large base measure \( G_{w_{i-2}w_{i-1}w_{i}} \) as a prior basis, the topic-dependent trigram is drawn by \( \text{PY} \) process \( H_{d,w_{i-2}w_{i-1}w_{i}} \sim \text{PY}(\theta_3, H_{d,w_{i-2}}) \). Therefore, HPYD process is recursively realized by sampling dependent topic n-gram probability \( p(w_i|w_{i-1}, w_{i-2}) \) and then n-gram probability \( p(w_i|w_{i-1}) \).
with probability $\frac{\gamma_k}{c_{w_{i+1}}}$ for a new table, this customer either

draws an existing menu $k$ with the probability $\frac{m_k}{\gamma_k}$ or a new

dish with probability $\frac{m_k}{\gamma_k}$ or a new menu with probability $\frac{m_k}{\gamma_k}$. The number of tables drawing

menu $k$ in all restaurants $m_k$ is used. Different tables may

choose the same menu. After selecting a table with menu $k$, the

customer $x_i$ further selects either an ordered dish $l$ with proba-

bility $\frac{m_k}{\gamma_k}$ or a new menu with probability $\frac{m_k}{\gamma_k}$. The number of dishes $\gamma_{lw}$ is counted over different menus. Combining

with new dishes provides the approach to model smoothing.

3.3. Gibbs sampling for inference of HPYD-LM

HPYD-LM is inferred according to Gibbs sampling based on

this new Chinese restaurant franchise. First, the general topic

measure is implemented as a mixture model for $K$ menus (or
topics), i.e., $G_k \sim \sum_{k=1}^{K} \gamma_k \frac{m_k}{\gamma_k} + \frac{m_k}{\gamma_k}$, where $\gamma_k$ is the atom of topic mixture model for menu $k$. Next,

draw topic-dependent unigram $H_{\theta(z \mid w)}$ for a word $w$ by

considering $G_k$ as a base measure according to the PY process.

With the prior measure $H_{\theta(z \mid w)}$, we draw unigram $G_{lw}$ by a

DP. In this fashion, HPYD $n$-gram with context $U = w_{i-n,+1}$

is recursively sampled by

\[
H_{w_{i-n+1}} \sim \frac{n_{kw_i}}{c_{w_{i+1}}} - \frac{m_k}{\gamma_k} \frac{\gamma_k}{c_{w_{i+1}}} + \frac{\gamma_k}{c_{w_{i+1}}} + \frac{\gamma_k}{c_{w_{i+1}}} + \frac{\gamma_k}{c_{w_{i+1}}} \frac{\gamma_k}{c_{w_{i+1}}} + \frac{\gamma_k}{c_{w_{i+1}}}
\]

which is realized from (4). In (5), the topic-dependent $n$-gram

$H_{w_{i-n+1}}$ is first drawn by a PY process with a prior

$G_{w_{i-n+1}}$ from backoff context $\pi(U) = w_{i-n+1}$, and subse-

quently treated as a prior to draw $n$-gram $G_{w_{i-n+1}}$ through

a DP. Using this HPYD-LM, backoff weights depend on the

number of dishes $\lambda_{kw_i}$ labelled by word $w_i$ in a menu $k$. The

topic-dependent $n$-grams are determined through drawing the

dishes from different contexts. Topic proportion is decided by

the number of customers $c_{w \mid \theta}(k)$ sitting in the tables which or-

der the same menu $k$. Latent topics are autonomously produced

by choosing new menus. Therefore, considering the topic-based

LM in (1) and the sitting and ordering arrangements of tables,

menus and dishes, we infer the nonparametric HPYD-LM

$p(w_i | w_{i-n+1})$ which is proportional to

\[
K \sum_{k=1}^{K} \frac{c_{w \mid \theta}(k)}{c_{w_{i+1}}} \left[ \frac{n_{kw_i}}{c_{w_{i+1}}} - \frac{m_k}{\gamma_k} \frac{\gamma_k}{c_{w_{i+1}}} + \frac{\gamma_k}{c_{w_{i+1}}} + \frac{\gamma_k}{c_{w_{i+1}}} \right]
\]

where $w = \{w_i, w_{i-1}, \ldots, t = \{t_i, t_{i-1}, \ldots, z = \{z_i, z_{i-1}, \ldots, \}$ and “-”

denotes the self-exclusion. The sitting arrangement is deter-

mined by sampling table $l$ according to either $p(l_0 | t_0, w, U)$

which is proportional to $c_{l_0}^T \cdot p(w_i | t_0, z, w, \ldots, U)$ if table $l$ is

occupied or $\gamma_n \cdot p(w_i | l_i = new, t_i, z, w, \ldots, U)$ if table $l$ is new. After sitting in a new table, we sample a distinct menu or topic

for this table given context $U$ by $p(z_i = k | x_i, t, w, \ldots, U)$ which is proportional to either $m_k \cdot p(w_i | l_i = new, t_i, z_i, w, \ldots, U)$ if menu $k$ is

ordered or $\gamma_n \cdot p(w_i | l_i = new, t_i, z_i, w, \ldots, U)$ if menu $k$ is new. Next, we draw a dish in menu $k$ by $p(l_i | z_i = k, z_{i-1}, w, \ldots, U)$ which is proportional to either $m_k \cdot p(w_i | l_i = new, t_i, z_i, w, \ldots, U)$ if dish $l_i = w$ is new. The counts $c_{wl}$ and $n_{lw}$ are measured

over all words except $w_i$. The

4. Experiments

4.1. Experimental setup

We evaluate the proposed HPYD-LM by using three datasets with
different contents and data sizes. The metrics of perplexity

and word error rate (WER) (%) are evaluated. In contin-

uous speech recognition, we adopted the Wall Street Journal

(WSJ) 1987-1989 corpus containing 86k documents with 38M

words and a vocabulary size of 5000. A total of 330 test sen-
tences were sampled from November 1992 ARPA CSR benchmark

data. The SI-84 training set was used to estimate HMM

parameters based on 39-dimensional MFCC feature vectors.

System configuration was detailed in [8]. Two other datasets

were collected for evaluation of perplexity. First, the Asso-

ciated Press newswire (AP) 1989 dataset consisted of 84,778

documents and 1,768,742 sentences with a vocabulary size of

1600 words. AP was partitioned into a training set with 36,727,591

words and a test set with 4,022,423 words. Second, NIPS0-12

(http://arabyn.net/resources.html) contained 1740 papers from

NIPS conferences. We collected a total of 2,034,215 words with

a vocabulary size of 3360. NIPS papers were divided into a

training set with 1,830,392 words and a test set with 203,823

words.

For comparative study, we carried out trigram LMs by using

LDA-LM [20], cDC-LM [8], MKN-LM [6][12] and HPY-LM

[11][22]. MKN-LM was carried out by using [18]. The results

of LDA-LM and cDC-LM were obtained by interpolating with

MKN-LM. HPY-LM and HPY-LM were implemented by per-

forming Gibbs sampling with 200 iterations and 100 samples at
each iteration. The burn-in samples in the first 20 iterations
were abandoned. Representative samples from the stationary

distribution were collected. The parameters $d_U \sim \beta(2,5)$

and $\theta_U \sim \text{Gam}(\alpha, \beta)$ were drawn with randomly selected

$\alpha$ and $\beta$. The parameter $\gamma_U = 100$ was fixed for all contexts

$U$ and $n$-grams. Perplexity of test data was examined by using

SRI toolkit [18] for comparison. We found that the larger the

number of words choosing a specific topic, the higher the aver-
gage discount based on $\lambda_{kw}$ is calculated. This phenomenon

meets the power law property.

4.2. Experimental results

First of all, Figure 2 displays the estimated topic proportions

(number of customers in the tables) in different topics (menus)
at different sampling iterations. WSJ corpus is used. The aver-
age log probability and the number of estimated topics are

shown. Gibbs sampling converges in these iterations. At each

iteration, the topics are adaptively estimated. Only a small num-

ber of topics have large topic proportions while many topics
have small topic proportions. Distribution of topic proportions conforms with power-law property. Table 1 shows an example of topic words from five selected topics which are extracted via HPYD-LM using WSJ corpus. It is obvious that topic words within a latent topic are semantically similar while those across topics are significantly different. This topic information is beneficial for estimation of language model. Using HPYD-LM, topics are automatically generated from data. The number of topics is affected by the seed parameter $\gamma_0$ as well as the number of n-gram events in training corpus. Larger $\gamma_0$ is more likely to draw new menus or lead to a larger topic model. Here, we empirically selected $\gamma_0 = 10$ for three datasets. Using this $\gamma_0$, the averaged number of topics in sampling process is 60 for WSJ, 55 for AP and 32 for NIPS. BNP learning works for scalable LMs under different size of training data.

Figure 3 displays the perplexity versus the number of topics or classes by using LDA-LM, cDC-LM, HPY-LM and the proposed HPYD-LM. WSJ corpus is adopted. Using LDA-LM and cDC-LM, the *parametric topic priors* are introduced in Bayesian topic-based LMs where the number of topics is fixed to be $K=20, 50, 100, 150$ and 200. We can see that perplexity is reduced by increasing number of topics. The lowest perplexity of LDA-LM (114.10) is achieved by using 150 topics. However, BNP learning based on HPY-LM and HPYD-LM reduces the perplexity as 112.20 and 106.84, respectively. Using HPYD-LM, the number of topics (60) is automatically determined. *Nonparametric topic priors* are introduced to build an effective and compact topic-based LM. Furthermore, Table 2 lists the perplexities of MKN-LM, LDA-LM, cDC-LM, HPY-LM and HPYD-LM by using AP and NIPS datasets. HPYD-LM achieves the lowest perplexity among these methods. In case of AP dataset, MKN-LM, LDA-LM, cDC-LM, HPY-LM and HPYD-LM obtain the perplexity of 115.82, 112.35, 109.11, 110.07 and 97.81, respectively. On the other hand, we investigate the performance of continuous speech recognition by using different LMs as shown in Table 3. In this comparison, WERs are reported under comparable model complexity. The results of LDA-LM and cDC-LM were implemented by fixing number of topics or classes to be 60. In this table, HPYD-LM and HPYD-LM2 imply the HPYD-LM with hyperparameters $\gamma_0 = 1$ and $\gamma_0 = 10$ and lead to 52 and 60 topics in the estimated HPYD-LMs, respectively. In this set of experiments, LDA-LM and cDC-LM are interpolated with MKN-LM, CDC-LM, HPY-LM and HPYD-LM. Among these LMs, the lowest WER 4.82% is achieved by using HPYD-LM with $\gamma_0 = 10$. HPYD-LM could estimate compact LM for speech recognition. The improvement using HPYD-LM comes from the flexibility of backoff smoothing and the contribution of scalable topic information.

<table>
<thead>
<tr>
<th>Trade</th>
<th>Investment</th>
<th>Stock Market</th>
<th>Politics</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>company</td>
<td>previously</td>
<td>talk</td>
<td>offer</td>
<td>party</td>
</tr>
</tbody>
</table>

**Table 1: Topic words using HPYD-LM under different topics.**

![Topic Proportion Graph](image1)

**Figure 2: Topic proportions in different sampling iterations.**

![Perplexity Graph](image2)

**Figure 3: Perplexity versus number of topics using WSJ dataset.**

![Perplexity Graph](image3)

**Figure 3: Perplexity versus number of topics using WSJ dataset.**

5. Conclusions

We presented the HPYD-LM based on a new random process which combined HPY for constructing the topic-dependent backoff smoothed LMs and HDP for integrating these LMs into a topic mixture model. Model selection issue was tackled by flexibly extending the number of topics. A Chinese restaurant franchise was proposed to implement the HPYD-LM which satisfied the properties for power-law distribution and topic mixture distribution. The posterior probabilities for drawing tables, menus and dishes were derived. Gibbs sampling was applied to infer HPYD-LM parameters. The experiments on WSJ, AP and NIPS showed that HPYD-LM outperformed the other LMs in terms of perplexity and word error rate.
6. References


