Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models

Kei Uchiumi **Hiroshi Tsukahara**
Denso IT Laboratory, Inc.
Shibuya Cross Tower 28F
2-15-1 Shibuya, Tokyo, Japan
{kuchiumi,htsukahara}@d-itlab.co.jp

Daichi Mochihashi
The Institute of Statistical Mathematics
10-3 Midori-cho, Tachikawa city
Tokyo, Japan
daichi@ism.ac.jp

Abstract

We propose a nonparametric Bayesian model for joint unsupervised word segmentation and part-of-speech tagging from raw strings. Extending a previous model for word segmentation, our model is called a Pitman-Yor Hidden Semi-Markov Model (PYHSMM) and considered as a method to build a class n-gram language model directly from strings, while integrating character and word level information. Experimental results on standard datasets on Japanese, Chinese and Thai revealed it outperforms previous results to yield the state-of-the-art accuracies. This model will also serve to analyze a structure of a language whose words are not identified a priori.

1 Introduction

Morphological analysis is a staple of natural language processing for broad languages. Especially for some East Asian languages such as Japanese, Chinese or Thai, word boundaries are not explicitly written, thus morphological analysis is a crucial first step for further processing. Note that also in Latin and old English, scripts were originally written with no word indications (*scripta continua*), but people felt no difficulty reading them. Here, morphological analysis means word segmentation and part-of-speech (POS) tagging.

For this purpose, supervised methods have often been employed for training. However, to train such supervised classifiers, we have to prepare a large amount of training data with correct annotations, in this case, word segmentation and POS tags. Creating and maintaining these data is not only costly but also very difficult, because generally there are no clear criteria for either "correct" segmentation or POS tags. In fact, since there are different standards for Chinese word segmentation, widely used SIGHAN Bakeoff dataset (Emerson, 2005) consists of multiple parts employing different annotation schemes.

Lately, this situation has become increasingly important because there are strong demands for processing huge amounts of text in consumer generated media such as Twitter, Weibo or Facebook (Figure 1). They contain a plethora of colloquial expressions and newly coined words, including sentiment expressions such as emoticons that cannot be covered by fixed supervised data.

To automatically recognize such linguistic phenomena beyond small "correct" supervised data, we have to extract linguistic knowledge from the statistics of strings themselves in an unsupervised fashion. Needless to say, such methods will also contribute to analyzing speech transcripts, classic texts, or even unknown languages. From a scientific point of view, it is worth while to find "words" and their part-of-speech purely from a collection of strings without any preconceived assumptions.

To achieve that goal, there have been two kinds of approaches: heuristic methods and statistical generative models. Heuristic methods are based on basic observations such that word boundaries will often occur at the place where predictive entropy of characters is large (i.e. the next character cannot be predicted without assuming

ローラのときに涙かブハァってなりました∩（´;ヮ;`）∩〜〜
真樹なんてこんな中２くさい事胸張って言えるぞぉ！
今日ね！らんらんとるいとコラボキャスするからおいで〜(⋆´∀`）ノシ
どうせ明日の昼ごろしれっと不在表入ってるんだろうなぁ。
テレ東はいつものネトウヨホルホル VTR 鑑賞番組してんのか

Figure 1: Sample of Japanese Twitter text that is difficult to analyze by ordinary supervised segmentation. It contains a lot of novel words, emoticons, and colloquial expressions.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing, pages 1774–1782,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

the next word). By formulating such ideas as search or MDL problems of given coding length[1], word boundaries are found in an algorithmic fashion (Zhikov et al., 2010; Magistry and Sagot, 2013). However, such methods have difficulty incorporating higher-order statistics beyond simple heuristics, such as word transitions, word spelling formation, or word length distribution. Moreover, they usually depends on tuning parameters like thresholds that cannot be learned without human intervention.

In contrast, statistical models are ready to incorporate all such phenomena within a consistent statistical generative model of a string, and often prove to work better than heuristic methods (Goldwater et al., 2006; Mochihashi et al., 2009). In fact, the statistical methods often include the criteria of heuristic methods at least in a conceptual level, which is noted in (Mochihashi et al., 2009) and also explained later in this paper. In a statistical model, each word segmentation $\mathbf{w}$ of a string s is regarded as a hidden stochastic variable, and the unsupervised learning of word segmentation is formulated as a maximization of a probability of $\mathbf{w}$ given s:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w}|s)\,. \qquad (1)$$

This means that we want the most "natural" segmentation $\mathbf{w}$ that have a high probability in a language model $p(\mathbf{w}|s)$.

Lately, Chen et al. (2014) proposed an intermediate model between heuristic and statistical models as a product of character and word HMMs. However, these two models do not have information shared between the models, which is not the case with generative models.

So far, these approaches only find word segmentation, leaving part-of-speech information behind. These two problems are not actually independent but interrelated, because knowing the part-of-speech of some infrequent or unknown word will give contextual clues to word segmentation, and vice versa. For example, in Japanese

すもももももも

can be segmented into not only すもも/も/もも/も (plum/too/peach/too), but also into すもも/もも/もも (plum/peach/peach), which is ungrammatical. However, we could exclude the latter case if we leverage knowledge that a state sequence N/P/N/P is much more plausible in Japanese than N/N/N from the part-of-speech information. Sirts and Alumäe (2012) treats a similar problem of POS induction with unsupervised morphological segmentation, but they know the words in advance and only consider segmentation within a word.

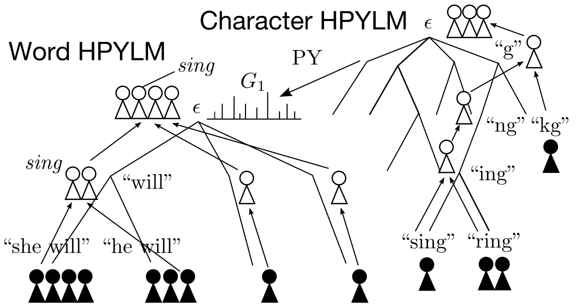

Figure 2: NPYLM represented in a hierarchical Chinese restaurant process. Here, a character ∞-gram HPYLM is embedded in a word n-gram HPYLM and learned jointly during inference.

For this objective, we attempt to maximize the joint probability of words and tags:

$$\underset{\mathbf{w},\mathbf{z}}{\operatorname{argmax}}\; p(\mathbf{w},\mathbf{z}|s) \propto p(\mathbf{w},\mathbf{z},s) \qquad (2)$$

From the expression above, this amounts to building a generative model of a string s with words $\mathbf{w}$ and tags $\mathbf{z}$ along with an associated inference procedure. We solve this problem by extending previous generative model of word segmentation. Note that heuristic methods are never able to model the hidden tags, and only statistical generative models can accommodate this objective.

This paper is organized as follows. In Section 2, we briefly introduce NPYLM (Mochihashi et al., 2009) on which our extension is based. Section 3 extends it to include hidden states to yield a hidden semi-Markov models (Murphy, 2002), and we describe its inference procedure in Section 4. We conduct experiments on some East Asian languages in Section 5. Section 6 discusses implications of our model and related work, and Section 7 concludes the paper.

2 Nested Pitman-Yor Language Model

Our joint model of words and states is an extension of the Nested Pitman-Yor Language Model (Mochihashi et al., 2009) of a string, which in turn is an extension of a Bayesian n-gram language model called Hierarchical Pitman-Yor Language Model (HPYLM) (Teh, 2006).

[1] For example, Zhikov et al. (2010) defined a coding length using character n-grams plus MDL penalty. Since this can be interpreted as a crude "likelihood" and a prior, its essence is similar but driven by a quite simplistic model.

HPYLM is a nonparametric Bayesian model of n-gram distribution based on the Pitman-Yor process (Pitman and Yor, 1997) that generates a discrete distribution G as $G \sim \mathrm{PY}(G_0, d, \theta)$. Here, d is a discount factor, "parent" distribution G_0 is called a base measure and θ controls how similar G is to G_0 in expectation. In HPYLM, n-gram distribution $G_n = \{p(w_t|w_{t-1} \cdots w_{t-(n-1)})\}$ is assumed to be generated from the Pitman-Yor process

$$G_n \sim \mathrm{PY}(G_{n-1}, d_n, \theta_n)\,, \tag{3}$$

where the base measure G_{n-1} is an $(n-1)$-gram distribution generated recursively in accordance with (3). Note that there are different G_n for each n-gram history $h = w_{t-1} \cdots w_{t-(n-1)}$. When we reach the unigram G_1 and need to use a base measure G_0, i.e. prior probabilities of words, HPYLM usually uses a uniform distribution over the lexicon.

However, in the case of unsupervised word segmentation, every sequence of characters could be a word, thus the size of the lexicon is unbounded. Moreover, prior probability of forming a word should not be uniform over all sequences of characters: for example, English words rarely begin with 'gme' but tend to end with '-ent' like in *segment*. To model this property, NPYLM assumes that word prior G_0 is generated from character HPYLM to model a well-formedness of w. In practice, to avoid dependency on n in the character model, we used an ∞-gram VPYLM (Mochihashi and Sumita, 2008) in this research. Finally, NPYLM gives an n-gram probability of word w given a history h recursively by integrating out G_n,

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_{h\cdot}}{\theta + c(h)}\, p(w|h')\,, \tag{4}$$

where h' is the shorter history of $(n-1)$-grams. $c(w|h), c(h) = \sum_w c(w|h)$ are n-gram counts of w appearing after h, and $t_{hw}, t_{h\cdot} = \sum_w t_{hw}$ are associated latent variables explained below. In case the history h is already empty at the unigram, $p(w|h') = p_0(w)$ is computed from the character ∞-grams for the word $w = c_1 \cdots c_k$:

$$p_0(w) = p(c_1 \cdots c_k) \tag{5}$$

$$= \textstyle\prod_{i=1}^{k} p(c_i|c_{i-1} \cdots c_1)\,. \tag{6}$$

In practice, we further corrected (6) so that a word length follows a mixture of Poisson distributions. For details, see (Mochihashi et al., 2009).

When we know word segmentation $\mathbf{w}$ of the data, the probability above can be computed by adding each n-gram count of w given h to the model, i.e. increment $c(w|h)$ in accordance with a hierarchical Chinese restaurant process associated with HPYLM (Figure 2). When each n-gram count called a customer is inferred to be actually generated from $(n-1)$-grams, we send its proxy customer for smoothing to the parent restaurant and increment t_{hw}, and this process will recurse. Notice that if a word w is never seen in $\mathbf{w}$, its proxy customer is eventually sent to the parent restaurant of unigrams. In that case[2], w is decomposed to its character sequence $c_1 \cdots c_k$ and this is added to the character HPYLM in the same way, making it a little "clever" about possible word spellings.

Inference Because we do not know word segmentation $\mathbf{w}$ beforehand, we begin with a trivial segmentation in which every sentence is a single word[3]. Then, we iteratively refine it by sampling a new word segmentation $\mathbf{w}(s)$ of a sentence s in a Markov Chain Monte Carlo (MCMC) framework using a dynamic programming, as is done with PCFG by (Johnson et al., 2007) shown in Figure 3 where we omit MH steps for computational reasons. Further note that every hyperparameter d_n, θ_n of NPYLM can be sampled from the posterior in a Bayesian fashion, as opposed to heuristic methods that rely on a development set for tuning. For details, see Teh (2006).

3 Pitman-Yor Hidden Semi-Markov Models

NPYLM is a complete generative model of a string, that is, a hierarchical Bayesian n-gram lan-

```
Input: a collection of strings $S$
Add initial segmentation $\mathbf{w}(s)$ to $\Theta$
for $j = 1 \cdots J$ do
  for $s$ in randperm ($S$) do
    Remove customers of $\mathbf{w}(s)$ from $\Theta$
    Sample w(s) according to $p(\mathbf{w}|s, \Theta)$
    Add customers of w(s) to $\Theta$
  end for
  Sample hyperparameters of $\Theta$
end for
```

Figure 3: MCMC inference of NPYLM Θ.

[2] To be precise, this occurs whenever t_{hw} is incremented in the unigram restaurant.

[3] Note that a child first memorizes what his mother says as a single word and gradually learns the lexicon.

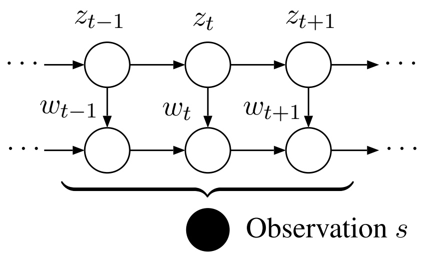

Figure 4: Graphical model of PYHSMM in a bigram case. White nodes are latent variables, and the shaded node is the observation. We only observe a string s that is a concatenation of hidden words $w_1 \cdots w_T$.

guage model combining words and characters. It can also be viewed as a way to build a Bayesian word n-gram language model directly from a sequence of characters, without knowing “words” a priori.

One possible drawback of it is a lack of part-of-speech: as described in the introduction, grammatical states will contribute much to word segmentation. Also, from a computational linguistics point of view, it is desirable to induce not only words from strings but also their part-of-speech purely from the usage statistics (imagine applying it to an unknown language or colloquial expressions). In classical terms, it amounts to building a class n-gram language model where both class and words are unknown to us. Is this really possible?

Yes, we can say it is possible. The idea is simple: we augment the latent states to include a hidden part-of-speech z_t for each word w_t, which is again unknown as displayed in Figure 4. Assuming w_t is generated from z_t’-th NPYLM, we can draw a generative model of a string s as follows:

```
$z_0 =$ BOS; $s = \epsilon$ (an empty string).
for $t = 1 \cdots T$ do
    Draw $z_t \sim p(z_t|z_{t-1})$ ,
    Draw $w_t \sim p(w_t|w_1 \cdots w_{t-1}, z_t)$ ,
    Append $w_t$ to $s$ .
end for
```

Here, z_0 = BOS and z_{T+1} = EOS are distinguished states for beginning and end of a sentence, respectively. For the transition probability of hidden states, we put a HPY process prior as (Blunsom and Cohn, 2011):

$$p(z_t|z_{t-1}) \sim \mathrm{HPY}(d, \theta) \tag{7}$$

with the final base measure being a uniform distribution over the states. The word boundaries are known in (Blunsom and Cohn, 2011), but in our case it is also learned from data at the same time. Note that because w_t depends on already generated words $w_1 \cdots w_{t-1}$, our model is considered as an autoregressive HMM rather than a vanilla HMM, as shown in Figure 4 ($w_{t-1} \rightarrow w_t$ dependency).

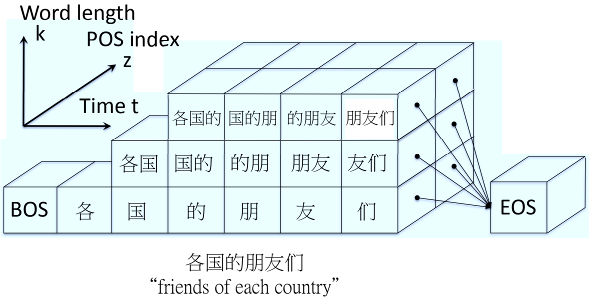

Figure 5: Graphical representation of sampling words and POSs. Each cell corresponds to an inside probability $\alpha[t][k][z]$. Note each cell is not always connected to adjacent cells, because of an overlap of substrings associated with each cell.

Since segment models like NPYLM have segment lengths as hidden states, they are called semi-Markov models (Murphy, 2002). In contrast, our model also has hidden part-of-speech, thus we call it a Pitman-Yor Hidden Semi-Markov model (PYHSMM).[4] Note that this is considered as a generative counterpart of a discriminative model known as a hidden semi-Markov CRF (Sarawagi and Cohen, 2005).

4 Inference

Inference of PYHSMM proceeds in almost the same way as NPYLM in Figure 3: For each sentence, first remove the customers associated with the old segmentation similarly to adding them. After sampling a new segmentation and states, the model is updated by adding new customers in accordance with the new segmentation and hidden states.

4.1 Sampling words and states

To sample words and states (part-of-speech) jointly, we first compute inside probabilities forward from BOS to EOS and sample backwards from EOS according to the Forward filtering-Backward sampling algorithm (Scott, 2002). This

[4] Lately, Johnson et al. (2013) proposed a nonparametric Bayesian hidden semi-Markov models for general state spaces. However, it depends on a separate distribution for a state duration, thus is clealy different from ours for a natural language.

can be regarded as a "stochastic Viterbi" algorithm that has the advantage of not being trapped in local minima, since it is a valid move of a Gibbs sampler in a Bayesian model.

For a word bigram case for simplicity, inside variable $\alpha[t][k][z]$ is a probability that a substring $c_1 \cdots c_t$ of a string $s = c_1 \cdots c_N$ is generated with its last k characters being a word, generated from state z as shown in Figure 5. From the definition of PYHSMM, this can be computed recursively as follows:

$$\alpha[t][k][z] = \sum_{j=1}^{L} \sum_{y=1}^{K} p(c_{t-k}^{t} | c_{t-k-j+1}^{t-k}, z) \, p(z|y) \alpha[t-k][j][y] \,. \quad (8)$$

Here, c_s^t is a substring $c_s \cdots c_t$ and L $(\leq t)$ is the maximum length of a word, and K is the number of hidden states.[5]

In Figure 5, each cell represents $\alpha[t][k][z]$ and a single path connecting from EOS to BOS corresponds to a word sequence $\mathbf{w}$ and its state sequence $\mathbf{z}$. Note that each cell is not always connected to adjacent cells (we omit the arrows), because the length-k substring associated with each cell already subsumes that of neighborhood cells.

Once $\mathbf{w}$ and $\mathbf{z}$ are sampled, each w_t is added to z_t'-th NPYLM to update its statistics.

4.2 Efficient computation by the Negative Binomial generalized linear model

Inference algorithm of PYHSMM has a computational complexity of $O(K^2 L^2 N)$, where N is a length of the string to analyze. To reduce computations it is effective to put a small L of maximum word length, but it might also ignore occasionally long words. Since these long words are often predictable from some character level information including suffixes or character types, in a semi-supervised setting we employ a Negative Binomial generalized linear model (GLM) for setting L_t adaptively for each character position t in the corpus.

Type	Feature
c_i	Character at time $t-i$ $(0 \leq i \leq 1)$
t_i	Character type at time $t-i$ $(0 \leq i \leq 4)$
cont	# of the same character types before t
ch	# of times character types changed within 8 characters before t

Table 1: Features used for the Negative Binomial generalized linear model for maximum word length prediction.

Specifically, we model the word length ℓ by a Negative Binomial distribution (Cook, 2009):

$$\ell \sim \mathrm{NB}(\ell | r, p) = \frac{\Gamma(r+\ell)}{\Gamma(r)\,\ell!} p^{\ell} (1-p)^r \,. \quad (9)$$

This counts the number of failures of Bernoulli draws with probability $(1-p)$ before r'th success. For our model, note that Negative Binomial is obtained from a Poisson distribution $\mathrm{Po}(\lambda)$ whose parameter λ again follows a Gamma distribution $\mathrm{Ga}(r, b)$ and integrated out:

$$p(\ell | r, b) = \int \mathrm{Po}(\ell | \lambda) \mathrm{Ga}(\lambda | r, b) d\lambda \quad (10)$$

$$= \frac{\Gamma(r+\ell)}{\Gamma(r)\,\ell!} \left(\frac{b}{1+b} \right)^{\ell} \left(\frac{1}{1+b} \right)^{r} . \quad (11)$$

This construction exactly mirrors the Poisson-Gamma word length distribution in (Mochihashi et al., 2009) with sampled λ. Therefore, our Negative Binomial is basically a continuous analogue of the word length distribution in NPYLM.[6]

Since $r > 0$ and $0 \leq p \leq 1$, we employ an exponential and sigmoidal linear regression

$$r = \exp(\mathbf{w}_r^T \mathbf{f}), \;\; p = \sigma(\mathbf{w}_p^T \mathbf{f}) \quad (12)$$

where $\sigma(x)$ is a sigmoid function and $\mathbf{w}_r, \mathbf{w}_p$ are weight vectors to learn. $\mathbf{f}$ is a feature vector computed from the substring $c_1 \cdots c_t$, including $f_0 \equiv 1$ for a bias term. Table 1 shows the features we used for this Negative Binomial GLM. Since Negative Binomial GLM is not convex in $\mathbf{w}_r$ and $\mathbf{w}_p$, we endow a Normal prior $\mathrm{N}(0, \sigma^2 I)$ for them and used a random walk MCMC for inference.

Predicting L_t Once the model is obtained, we can set L_t adaptively as the time where the cumulative probability of ℓ exceeds some threshold θ (we used $\theta = 0.99$). Table 2 shows the precision of predicting maximum word length learned from 10,000 sentences from each set: it measures whether the correct word boundary in test data is included in the predicted L_t.

Overall it performs very well with high precision, and works better for longer words that cannot be accommodated with a fixed maximum length.

[5] For computational reasons, we do not pursue using a Dirichlet process to yield an infinite HMM (Van Gael et al., 2009), but it is straightforward to extend our PYHSMM to iHMM.

[6] Because NPYLM employs a mixture of Poisson distributions for each character type of a substring, this correspondence is not exact.

Lang	Dataset	Training	Test
Ja	Kyoto corpus	37,400	1,000
	BCCWJ OC	20,000	1,000
Zh	SIGHAN MSR	86,924	3,985
	SIGHAN CITYU	53,019	1,492
	SIGHAN PKU	19,056	1,945
Th	InterBEST Novel	1,000	1,000

Table 3: Datasets used for evaluation. Abbreviations: Ja=Japanese, Zh=Chinese, Th=Thai language.

Figure 6 shows the distribution of predicted maximum lengths for Japanese. Although we used $\theta = 0.99$, it is rather parsimonious but accurate that makes the computation faster.

Because this cumulative Negative Binomial prediction is language independent, we believe it might be beneficial for other natural language processing tasks that require some maximum lengths within which to process the data.

5 Experiments

To validate our model, we conducted experiments on several corpora of East Asian languages with no word boundaries.

Datasets For East Asian languages, we used standard datasets in Japanese, Chinese and Thai as shown in Table 3. The Kyoto corpus is a collection of sentences from Japanese newspaper (Kurohashi and Nagao, 1998) with both word segmentation and part-of-speech annotations. BC-CWJ (Balanced Corpus of Contemporary Written Japanese) is a balanced corpus of written Japanese (Maekawa, 2007) from the National Institute of Japanese Language and Linguistics, also with both word segmentation and part-of-speech annotations from slightly different criteria. For experiments on colloquial texts, we used a random subset of "OC" register from this corpus that is comprised of Yahoo!Japan Answers from users. For Chinese, experiments are conducted on standard datasets of SIGHAN Bakeoff 2005 (Emerson, 2005); for comparison we used MSR and PKU datasets for simplified Chinese, and the CITYU dataset for traditional Chinese. SIGHAN datasets have word boundaries only, and we conformed to original training/test splits provided with the data. InterBEST is a dataset in Thai used in the InterBEST 2009 word segmentation contest (Kosawat, 2009). For contrastive purposes, we used a "Novel" subset of it with a random sampling without replacement for training and test data. Accuracies are measured in token F-measures computed as follows:

$$F = \frac{2PR}{P+R}, \tag{13}$$

$$P = \frac{\textit{\# of correct words}}{\textit{\# of words in output}}, \tag{14}$$

$$R = \frac{\textit{\# of correct words}}{\textit{\# of words in gold standard}}. \tag{15}$$

Unsupervised word segmentation In Table 4, we show the accuracies of unsupervised word segmentation with previous figures. We used bigram PYHSMM and set $L = 4$ for Chinese, $L = 5, 8, 10, 21$ for Japanese with different types of contiguous characters, and $L = 6$ for Thai. The number of hidden states are $K = 10$ (Chinese and Thai), $K = 20$ (Kyoto) and $K = 30$ (BCCWJ).

We can see that our PYHSMM outperforms on all the datasets. Huang and Zhao (2007) reports that the maximum possible accuracy in unsupervised Chinese word segmentation is 84.8%, derived through the inconsistency between different segmentation standards of the SIGHAN dataset. Our PYHSMM performs nearer to this best possible accuracy, leveraging both word and character knowledge in a consistent Bayesian fashion. Further note that in Thai, quite high performance is achieved with a very small data compared to previous work.

Unsupervised part-of-speech induction As stated above, Kyoto, BCCWJ and Weibo datasets

Dataset	Kyoto	BCCWJ	MSR	CITYU	BEST
Precision (All)	99.9	99.9	99.6	99.9	99.0
Precision (≥ 5)	96.7	98.4	73.6	87.0	91.7
Maximum length	15	48	23	12	21

Table 2: Precision of maximum word length prediction with a Negative Binomial generalized linear model (in percent). ≥ 5 are figures for word length ≥ 5. Final row is the maximum length of a word found in each dataset.

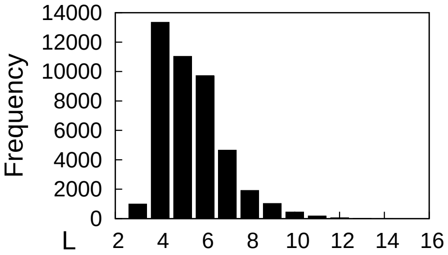

Figure 6: Distribution of predicted maximum word lengths on the Kyoto corpus.

Dataset	PYHSMM	NPY	BE	HMM^2
Kyoto	**71.5**	62.1	71.3	NA
BCCWJ	**70.5**	NA	NA	NA
MSR	**82.9**	80.2	78.2	81.7
CITYU	**82.6***	82.4	78.7	NA
PKU	**81.6**	NA	80.8	81.1
BEST	**82.1**	NA	**82.1**	NA

Table 4: Accuracies of unsupervised word segmentation. BE is a Branching Entropy method of Zhikov et al. (2010), and HMM^2 is a product of word and character HMMs of Chen et al. (2014). * is the accuracy decoded with $L=3$: it becomes 81.7 with $L=4$ as MSR and PKU.

have part-of-speech annotations as well. For these data, we also evaluated the precision of part-of-speech induction on the output of unsupervised word segmentation above. Note that the precision is measured only over correct word segmentation that the system has output. Table 5 shows the precisions; to the best of our knowledge, there are no previous work on joint unsupervised learning of words and tags, thus we only compared with Bayesian HMM (Goldwater and Griffiths, 2007) on both NPYLM segmentation and gold segmentation. In this evaluation, we associated each tag of supervised data with a latent state that cooccurred most frequently with that tag. We can see that the precision of joint POS tagging is better than NPYLM+HMM, and even better than HMM that is run over the gold segmentation.

For colloquial Chinese, we also conducted an experiment on the Leiden Weibo Corpus (LWC), a corpus of Chinese equivalent of Twitter[7]. We used random 20,000 sentences from this corpus, and results are shown in Figure 7. In many cases plausible words are found, and assigned to syntactically consistent states. States that are not shown here are either just not used or consists of a mixture of different syntactic categories. Guiding our model to induce more accurate latent states is a common problem to all unsupervised part-of-speech induction, but we show some semi-supervised results next.

Dataset	PYHSMM	NPY+HMM	HMM
Kyoto	**57.4**	53.8	49.5
BCCWJ	**50.2**	44.1	44.2
LWC	33.0	30.9	32.9

Table 5: Precision of POS tagging on correctly segmented words.

Semi-supervised experiments Because our PYHSMM is a generative model, it is easily amenable to semi-supervised segmentation and tagging. We used random 10,000 sentences from supervised data on Kyoto, BCCWJ, and LWC datasets along with unsupervised datasets in Table 3.

Results are shown in Table 6: segmentation accuracies came close to 90% but do not go beyond. By inspecting the segmentation and POS that PYHSMM has output, we found that this is not necessarily a fault of our model, but it came from the often inconsistet or incorrect tagging of the dataset. In many cases PYHSMM found more "natural" segmentations, but it does not always conform to the gold annotations. On the other hand, it often oversegments emotional expressions (sequence of the same character, for example) and this is one of the major sources of errors.

Finally, we note that our proposed model for unsupervised learning is most effective for the language which we do not know its syntactic behavior but only know raw strings as its data. In Figure 8, we show an excerpt of results to model a Japanese local dialect (*Mikawa-ben* around Nagoya district) collected from a specific Twitter. Even from the surface appearance of characters, we can see that similar words are assigned to the same state including some emoticons (states 9,29,32), and in fact we can identify a state of postpositions specific to that dialect (state 3). Notice that the words themselves are not trivial before this analysis. There are also some name of local places (state 41) and general Japanese postpositions (2) or nouns (11,18,25,27,31). Because of the sparsity promoting prior (7) over the hidden states, actually used states are sparse and the results can be considered quite satisfactory.

6 Discussion

The characteristics of NPYLM is a Baysian integration of character and word level information, which is related to (Blunsom and Cohn, 2011) and the adaptor idea of (Goldwater et al., 2011). This

Dataset	Seg	POS
Kyoto	92.1	87.1
BCCWJ	89.4	83.1
LWC	88.5	86.9

Table 6: Semi-supervised segmentation and POS tagging accuracies. POS is measured by precision.

[7] http://lwc.daanvanesch.nl/

$z=1$		$z=3$		$z=10$		$z=11$		$z=18$	
啦	227	。	3309	，	13440	可以	207	东	68
呀	182	！	1901	＃	5989	呢	201	大	60
去	86	了	482	的	5224	。	199	南	59
开心	65	啊	226	。	3237	那么	192	，	55
走	62	呢	110	我	1504	多	192	西	53
哈	53	哦	93	是	1206	打	177	路	51
鸟	44	啦	69	！	1190	才	167	海	49
喽	41	哈哈	56	在	900	比	165	山	49
波	31	地址	47	都	861	对	154	区	45
测试	30	晚安	43	和	742	几	146	去	39

Figure 7: Some interesting words and states induced from Weibo corpus ($K=20$). Numbers represent frequencies that each word is generated from that class. Although not perfect, emphatic ($z=1$), end-of-sentence expressions ($z=3$), and locative words ($z=18$) are learned from tweets. Distinction is far more clear in the semi-supervised experiments (not shown here).

z	Induced words
2	の 、 は に が で と も を 「
3	ぞん かん ね のん だに だん りん か ん だのん
9	（*^^*） ！（^−^; （^_^;） （^^;; ！（^^;;
10	。 ！ !! ？ 」 （≧▽≦） !! 」「
11	楽 入 ど寒 大丈夫 会 受 停電 良 美味 台風が
13	に ら わ な よ ね だら じゃんね え ぁ
18	今年 最近 豊川 地元 誰 豊田 今度 次 豊川高校
19	さん ん め 食べ って よろしく ありがとう じゃん
20	これ 知 人 それ どこ まあ みんな 東京 いや 方
24	三河弁 この よ お 何 そ ほい 今日 また ほ
25	他 一緒 5 大変 頭 春 参加 指 世代 地域
26	マジ 豊橋カレー コレ トキワ コーヒー プロ ファン
27	行 」 方言 & 言葉 普通 夜店 」 始 確認
29	（ ！（; （´・ !! (*` ?（´・ (*^_^*)
30	気 うち 店 ほう ここ こっち 先生 友人 いろいろ
31	女子 無理 決 近い 安心 標準語 感動 蒲郡 試合
32	（ （ *\(^ ＼（^ （^ ！*\(^ 〜 （^_^ （*^
34	ヤマサ マーラ オレ ハイジ イメージクッピーラムネ
35	な ー そう 好き こと らん なん ら み 意味
36	いい ど うまい 杏果 ぐろ めっちゃ かわい はよ
41	豊橋 名古屋 三河 西三河 名古屋弁 名古屋人 大阪

Figure 8: Unsupervised analysis of a Japanese local dialect by PYHSMM. ($K=50$)

is different from (and misunderstood in) a joint model of Chen et al. (2014), where word and character HMMs are just multiplied. There are no information shared from the model structure, and in fact it depends on a BIO-like heuristic tagging scheme in the character HMM.

In the present paper, we extended it to include a hidden state for each word. Therefore, it might be interesting to introduce a hidden state also for each character. Unlike western languages, there are many kinds of Chinese characters that work quite differently, and Japanese uses several distinct kinds of characters, such as a Chinese character, Hiragana, Katakana, whose mixture would constitute a single word. Therefore, statistical modeling of different types of characters is an important research venue for the future.

NPYLM has already applied and extended to speech recognition (Neubig et al., 2010), statistical machine translation (Nguyen et al., 2010), or even robotics (Nakamura et al., 2014). For all these research area, we believe PYHSMM would be beneficial for their extension.

7 Conclusion

In this paper, we proposed a Pitman-Yor Hidden Semi-Markov model for joint unsupervised word segmentation and part-of-speech tagging on a raw sequence of characters. It can also be viewed as a way to build a class n-gram language model directly on strings, without any “word” information a priori.

We applied our PYHSMM on several standard datasets on Japanese, Chinese and Thai, and it outperformed previous figures to yield the state-of-the-art results, as well as automatically induced word categories. It is especially beneficial for colloquial text, local languages or speech transcripts, where not only words themselves are unknown but their syntactic behavior is a focus of interest.

In order to adapt to human standards given in supervised data, it is important to conduct a semi-supervised learning with discriminative classifiers. Since semi-supervised learning requires generative models in advance, our proposed Bayesian generative model will also lay foundations to such an extension.

References

Phil Blunsom and Trevor Cohn. 2011. A Hierarchical Pitman-Yor Process HMM for Unsupervised Part of Speech Induction. In *ACL 2011*, pages 865–874.

Miaohong Chen, Baobao Chang, and Wenzhe Pei. 2014. A Joint Model for Unsupervised Chinese Word Segmentation. In *EMNLP 2014*, pages 854–863.

John D. Cook. 2009. Notes on the Negative Binomial Distribution. http://www.johndcook.com/negative_binomial.pdf.

Tom Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Sharon Goldwater and Tom Griffiths. 2007. A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. In *Proceedings of ACL 2007*, pages 744–751.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual Dependencies in Unsupervised Word Segmentation. In *Proceedings of ACL/COLING 2006*, pages 673–680.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2011. Producing Power-Law Distributions and Damping Word Frequencies with Two-Stage Language Models. *Journal of Machine Learning Research*, 12:2335–2382.

Chang-Ning Huang and Hai Zhao. 2007. Chinese word segmentation: A decade review. *Journal of Chinese Information Processing*, 21(3):8–20.

Matthew J. Johnson and Alan S. Willsky. 2013. Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research*, 14:673–701.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian Inference for PCFGs via Markov Chain Monte Carlo. In *Proceedings of HLT/NAACL 2007*, pages 139–146.

Krit Kosawat. 2009. InterBEST 2009: Thai Word Segmentation Workshop. In *Proceedings of 2009 Eighth International Symposium on Natural Language Processing (SNLP2009)*, Thailand.

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of LREC 1998*, pages 719–724. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html.

Kikuo Maekawa. 2007. Kotonoha and BCCWJ: Development of a Balanced Corpus of Contemporary Written Japanese. In *Corpora and Language Research: Proceedings of the First International Conference on Korean Language, Literature, and Culture*, pages 158–177.

Pierre Magistry and Benoît Sagot. 2013. Can MDL Improve Unsupervised Chinese Word Segmentation? In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 2–10.

Daichi Mochihashi and Eiichiro Sumita. 2008. The Infinite Markov Model. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1017–1024.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proceedings of ACL-IJCNLP 2009*, pages 100–108.

Kevin Murphy. 2002. Hidden semi-Markov models (segment models). http://www.cs.ubc.ca/˜murphyk/Papers/segment.pdf.

Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Shogo Nagasaka, Tadahiro Taniguchi, and Naoto Iwahashi. 2014. Mutual Learning of an Object Concept and Language Model Based on MLDA and NPYLM. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'14)*, pages 600–607.

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. 2010. Learning a Language Model from Continuous Speech. In *Proc. of INTERSPEECH 2010*.

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric Word Segmentation for Machine Translation. In *COLING 2010*, pages 815–823.

Jim Pitman and Marc Yor. 1997. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *Annals of Probability*, 25(2):855–900.

Sunita Sarawagi and William W. Cohen. 2005. Semi-Markov Conditional Random Fields for Information Extraction. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 1185–1192.

Steven L. Scott. 2002. Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association*, 97:337–351.

Kairit Sirts and Tanel Alumäe. 2012. A Hierarchical Dirichlet Process Model for Joint Part-of-Speech and Morphology Induction. In *NAACL 2012*, pages 407–416.

Yee Whye Teh. 2006. A Bayesian Interpretation of Interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, NUS.

Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *EMNLP 2009*, pages 678–687.

Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. 2010. An Efficient Algorithm for Unsupervised Word Segmentation with Branching Entropy and MDL. In *EMNLP 2010*, pages 832–842.