Non-parametric Bayesian Segmentation of Japanese Noun Phrases

Yugo Murawaki and **Sadao Kurohashi**
Graduate School of Informatics
Kyoto University
{murawaki, kuro}@i.kyoto-u.ac.jp

Abstract

A key factor of high quality word segmentation for Japanese is a high-coverage dictionary, but it is costly to manually build such a lexical resource. Although external lexical resources for human readers are potentially good knowledge sources, they have not been utilized due to differences in segmentation criteria. To supplement a morphological dictionary with these resources, we propose a new task of Japanese noun phrase segmentation. We apply non-parametric Bayesian language models to segment each noun phrase in these resources according to the statistical behavior of its supposed constituents in text. For inference, we propose a novel block sampling procedure named hybrid type-based sampling, which has the ability to directly escape a local optimum that is not too distant from the global optimum. Experiments show that the proposed method efficiently corrects the initial segmentation given by a morphological analyzer.

1 Introduction

Word segmentation is the first step of natural language processing for Japanese, Chinese and Thai because they do not delimit words by white-space. Segmentation for Japanese is a successful field of research, achieving the F-score of nearly 99% (Kudo et al., 2004). This success rests on a high-coverage dictionary. Unknown words, or words not covered by the dictionary, are often misidentified.

Historically, researchers have devoted extensive human resources to build and maintain high-coverage dictionaries (Yokoi, 1995). Since the orthography of Japanese does not specify a standard for segmentation, researchers define their own criteria before constructing lexical resources. For this reason, it is difficult to exploit existing external resources, such as dictionaries and encyclopedias for human readers, where entry words are not segmented according to the criteria. Among them, encyclopedias are especially important in that they contain a lot of terms that a morphological dictionary fails to cover. Most of these terms are noun phrases and consist of more than one word (morpheme). For example, an encyclopedia has an entry "常山城" (*tsuneyama-jou*, "Tsuneyama Castle"). According to our segmentation criteria, it consists of two words "常山" (*tsuneyama*) and "城" (*jou*). However, the morphological analyzer wrongly segments it into "常" (*tsune*) and "山城" (*yamashiro*) because "常山" (*tsuneyama*) is an unknown word.

In this paper, we present the first attempt to utilize encyclopedias for word segmentation. We segment each entry noun phrase into words. To do this, we examine the main text of the entry, on the assumption that if the noun phrase in question consists of more than one word, its constituents appear in the main text either freely or as part of other noun phrases. For "常山城" (*tsuneyama-jou*), its constituent "常山" (*tsune*) appears by itself and as constituents of other nouns phrases such as "常山山頂" (peak of Tsuneyama) and "常山駅" (Tsuneyama Station) while "山城" (*yamashiro*) does not.

To segment each noun phrase, we use non-parametric Bayesian language models (Goldwater et al., 2009; Mochihashi et al., 2009). Our approach

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 605–615,
Edinburgh, Scotland, UK, July 27–31, 2011.

is based on two key factors: the bigram model and type-based block sampling. The bigram model alleviates a problem of the unigram model, that is, a tendency to misidentify a sequence of words in common collocations as a single word. Type-based sampling (Liang et al., 2010) has the ability to directly escape a local optimum, making inference very efficient. However, type-based sampling is not easily applicable to the bigram model owing to sparsity and its dependence on latent assignments.

We propose a hybrid type-based sampling procedure, which combines the Metropolis-Hastings algorithm with Gibbs sampling. We circumvent the sparsity problem by joint sampling of unigram-level type. Also, instead of calculating the probability of every possible state of the jointly sampled random variables, we only compare the current state with a proposed state. This greatly eases the sampling procedure while retaining the efficiency of type-based sampling. Experiments show that the proposed method quickly corrects the initial segmentation given by a morphological analyzer.

2 Related Work

Japanese Morphological Analysis and Lexical Acquisition Word segmentation for Japanese is usually solved as the joint task of segmentation and part-of-speech tagging, which is called morphological analysis (Kurohashi et al., 1994; Asahara and Matsumoto, 2000; Kudo et al., 2004). The standard approach in Japanese morphological analysis is lattice-based path selection instead of character-based IOB tagging. Given a sentence, an analyzer first builds a lattice of words with dictionary look-up and then selects an optimal path using pre-defined parameters. This approach enables fast decoding and achieves accuracy high enough for practical use.

This success, however, depends on a high-coverage dictionary, and unknown words are often misidentified. Although a line of research attempts to identify unknown words on the fly (Uchimoto et al., 2001; Asahara and Matsumoto, 2004), it by no means provides a definitive solution because it suffers from locality of contextual information available for identification (Nakagawa and Matsumoto, 2006). Therefore we like to perform separate lexical acquisition processes in which wider context can be examined.

Our approach in this paper has a complementary relationship with unknown word acquisition from text, which we previously proposed (Murawaki and Kurohashi, 2008). Since, unlike Chinese and Thai, Japanese is rich in morphology, morphological regularity can be used to determine if an unknown word candidate in text is indeed the word to be acquired. In general, this method works pretty well, but one exception is noun phrases. Noun phrases can hardly be distinguished from single nouns because in Japanese, no morphological marker is attached to join nouns to form a noun phrase. We previously resort to a heuristic measure to segment noun phrases. The new statistical method provides a straightforward solution to this problem.

Meanwhile, our language models have their own problem. The assumption that language is a sequence of invariant words fails to capture rich morphology, as our segmentation criteria specify that each verb or adjective consists of an invariant stem and an ending that changes its form according to its grammatical roles. For this reason, we limit our scope to noun phrases in this paper.

Use of Noun Phrases Named entity recognition (NER) is a field where encyclopedic knowledge plays an important role. Kazama and Torisawa (2008) encode information extracted from a gazetteer (e.g. Wikipedia) as features of a CRF-based Japanese NE tagger. They formalize the NER task as the character-based labeling of IOB tags. Noun phrases extracted from a gazetteer are also straightforwardly represented as IOB tags. However, this does not fully solve the knowledge bottleneck problem. They also used the output of a morphological analyzer, which does not utilize encyclopedic knowledge. NER performance may be affected by segmentation errors in morphological analysis involving unknown words.

Chinese word segmentation is often formalized as a character tagging problem (Xue, 2003). In this setting, it is easy to incorporate external resources into the model. Low et al. (2005) introduce an external dictionary as features of a discriminative model. However, they only use words up to 4 characters in length. We conjecture that words in their dictionary are not noun phrases. External resources used by

Peng et al. (2004) are also lists of short words and characters.

Non-parametric Language Models Non-parametric Bayesian statistics offers an elegant solution to the task of unsupervised word segmentation, in which the vocabulary size is not known in advance (Goldwater et al., 2009; Mochihashi et al., 2009). It does not compete with supervised segmentation, however. Unsupervised word segmentation is used elsewhere, for example, with theoretical interest in children's language acquisition (Johnson, 2008; Johnson and Demuth, 2010) and with the application to statistical machine translation, in which segmented text is merely an intermediate representation (Xu et al., 2008; Nguyen et al., 2010). In this paper we demonstrate that non-parametric models can complement supervised segmentation.

3 Japanese Noun Phrase Segmentation

Our goal is to overcome the unknown word problem in morphological analysis by utilizing existing resources such as dictionaries and encyclopedias for human readers. In our settings, we are given a list of entries from external resources. Almost all of them are noun phrases and each entry consists of one or more words.

A naïve implementation would be to use noun phrases as they are. In fact, ipadic[1] regards as single words a large number of long proper nouns like "関西国際空港会社連絡橋" (literally, Kansai International Airport Company Connecting Bridge). However, this approach has various drawbacks. For example, in information retrieval, the query "Kansai International Airport" does not match the "single" word for the bridge. So we apply segmentation.

Each entry is associated with text, which is usually the main text of the entry.[2] We assume the text as the key to segmenting the noun phrase. If the noun phrase in question consists of more than one word, its constituents would appear in the text either freely or as part of other noun phrases.

We obtain the segmentation of an entry noun phrase by considering the segmentation of the whole text. One may instead consider a pipeline approach in which we first extract noun phrases in text and then identify boundaries within these noun phrases. However, noun phrases in text are not trivially identifiable in the case that they contain unknown words as their constituents. For example, the analyzer erroneously segments the word "ちんすこう" (*chiNsukou*) into "ちん" (*chiN*) and "すこう" (*sukou*), and since the latter is misidentified as a verb, the incorrect noun phrase "ちん" (*chiN*) is extracted.

We have a morphological analyzer with a dictionary that covers frequent words. Although it often misidentifies unknown words, the overall accuracy is reasonably high. For this reason, we like to use the segmentation given by the analyzer as the initial state and to make small changes to them to get a desired output. We also use an annotated corpus, which was used to build the analyzer. As the annotated corpus encodes our segmentation criteria, it can be used to force the models to stick with our segmentation criteria.

We concentrate on segmentation in this paper, but we also need to assign a POS tag to each constituent word and to incorporate segmented noun phrases into the dictionary of the morphological analyzer. We leave them for future work.[3]

4 Non-parametric Bayesian Language Models

To correct the initial segmentation given by the analyzer, we use non-parametric Bayesian language models that have been applied to unsupervised word segmentation (Goldwater et al., 2009). Specifically, we adopt unigram and bigram models. We propose a small modification to these models in order to exploit an annotated corpus when it is much larger than raw text.

4.1 Unigram Model

In the unigram model, a word in the corpus w_i is generated as follows:

$$\begin{aligned} G|\alpha_0, P_0 &\sim \mathrm{DP}(\alpha_0, P_0) \\ w_i|G &\sim G \end{aligned}$$

[1] http://sourceforge.jp/projects/ipadic/

[2] We may augment the text with related documents if the main text is not large enough.

[3] Fortunately, the morphological analyzer JUMAN is capable of handling *phrases*, each of which consists of more than one word. All we need to do is POS tagging.

where G is a distribution over a countably infinite set of words, and $\mathrm{DP}(\alpha_0, P_0)$ is a Dirichlet process (Ferguson, 1973) with the concentration parameter α_0 and the base distribution P_0, for which we use a zerogram model described in Section 4.3.

Marginalizing out G, we can interpret the model as a Chinese restaurant process. Suppose that we have observed $i-1$ words $\mathbf{w_{-i}} = w_1, \cdots, w_{i-1}$, the probability of w_i is given by

$$P_1(w_i = w|\mathbf{w_{-i}}) = \frac{n_w^{\mathbf{w_{-i}}} + \alpha_0 P_0}{i - 1 + \alpha_0}, \qquad (1)$$

where $n_w^{\mathbf{w_{-i}}}$ is the number of word label w observed in $\mathbf{w_{-i}}$.

The unigram model is known for its tendency to misidentify a sequence of words in common collocations as a single word (Goldwater et al., 2009). In preliminary experiments, we found that the unigram model often interpreted a noun phrase as a single word, even in the case that its constituents frequently appeared in text.

4.2 Bigram Model

The problem of the unigram model can be alleviated by the bigram model based on a hierarchical Dirichlet process (Goldwater et al., 2009). In the bigram model, word w_i is generated as follows:

$$\begin{aligned} G|\alpha_0, P_0 &\sim \mathrm{DP}(\alpha_0, P_0) \\ H_l|\alpha_1, G &\sim \mathrm{DP}(\alpha_1, G) \\ w_i|w_{i-1} = l, H_l &\sim H_l \end{aligned}$$

Marginalizing out G and H_l, we can again explain the model with the Chinese restaurant process. Unlike the unigram model, however, the bigram model depends on the latent table assignments $\mathbf{z_{-i}}$.

$$P_2(w_i|\mathbf{h_{-i}}) = \frac{n_{(w_{i-1},w_i)}^{\mathbf{h_{-i}}} + \alpha_1 P_1(w_i|\mathbf{h_{-i}})}{n_{(w_{i-1},*)}^{\mathbf{h_{-i}}} + \alpha_1} \qquad (2)$$

$$P_1(w_i|\mathbf{h_{-i}}) = \frac{t_{w_i}^{\mathbf{h_{-i}}} + \alpha_0 P_0(w_i)}{t_*^{\mathbf{h_{-i}}} + \alpha_0} \qquad (3)$$

where $\mathbf{h_{-i}} = (\mathbf{w_{-i}}, \mathbf{z_{-i}})$, $t_{w_i}^{\mathbf{h_{-i}}}$ is the number of tables labeled with w_i and $t_*^{\mathbf{h_{-i}}}$ is the total number of tables. Thanks to exchangeability, we do not need to track the exact seating assignments. Still, we need to maintain a *histogram* for each w that consists of frequencies of table customers (Blunsom et al., 2009).

4.3 Zerogram Model

Following Nagata (1996) and Mochihashi et al. (2009), we model the zerogram distribution P_0 with the word length k and the character sequence $w = c_1, \cdots, c_k$. Specifically, we define P_0 as the combination of a Poisson distribution with mean λ and a bigram distribution over characters.

$$\begin{aligned} P_0(w) &= P(k;\lambda)\frac{P(c_1, \cdots, c_k, k|\Theta)}{P(k|\Theta)} \\ P(k;\lambda) &= e^{-\lambda}\frac{\lambda^k}{k!} \\ P(c_1, \cdots, c_k, k|\Theta) &= \prod_{i=1}^{k+1} P(c_i|c_{i-1}) \end{aligned}$$

Θ is the zerogram model, and c_0 and c_{k+1} are a word boundary marker. $P(k|\Theta)$ can be estimated by randomly generating words from the model. We use different λ for different scripts. The Japanese writing system uses several scripts, and each word can be classified by script such as hiragana, katakana, kanji, the mixture of hiragana and kanji, etc. The optimal value for λ depends on scripts. For example, katakana, which predominantly denotes loan words, is longer on average than hiragana, which is often used for short function words.

We obtain the parameters and counts from an annotated corpus and fix them during noun phrase segmentation. This greatly simplifies inference but may make the model fragile with unknown words. For this reason, we set a hierarchical Pitman-Yor process prior (Teh, 2006; Goldwater et al., 2006) for the bigram probability $P(c_i|c_{i-1})$ with the base distribution of character unigrams. Note that even character bigrams are sparse because thousands of characters are used in Japanese.

4.4 Mixing an Annotated Corpus

An annotated corpus can be used to force the models to stick with our segmentation criteria. A straightforward way to do this is to mix it with raw text while fixing the segmentation during inference (Mochihashi et al., 2009). A word found in the annotated corpus is generally preferred because it has fixed counts obtained from the annotated corpus. We call this method direct mixing.

Direct mixing is problematic when raw text is much smaller than the annotated corpus. With this

situation, the role of raw text associated with the noun phrase in question is marginalized by the annotated corpus.

As a solution to this problem, we propose another mixing method called back-off mixing. In back-off mixing, the annotated corpus is used as part of the base distribution. In the unigram model, P_0 in (1) is replaced by

$$P_0^{\mathrm{BM}} = \lambda_{\mathrm{IP}} P_0 + (1 - \lambda_{\mathrm{IP}}) P_1^{\mathrm{REF}},$$

where λ_{IP} is a parameter for linear interpolation and P_1^{REF} is the unigram probability obtained from the annotated text. The loose coupling makes the models robust to an imbalanced pair of texts. Similarly, the back-off mixing bigram model replaces P_1 in (2) with

$$P_1^{\mathrm{BM}} = \lambda_{\mathrm{IP}} P_1 + (1 - \lambda_{\mathrm{IP}}) P_2^{\mathrm{REF}}.$$

5 Inference

Collapsed Gibbs sampling is widely used to find an optimal segmentation (Goldwater et al., 2009). In this section, we first show that simple collapsed sampling can hardly escape the initial segmentation. To address this problem, we apply a block sampling algorithm named type-based sampling (Liang et al., 2010) to the unigram model. Since type-based sampling is not applicable to the bigram model, we propose a novel sampling procedure for the bigram model, which we call hybrid type-based sampling.

5.1 Collapsed Sampling

In collapsed Gibbs sampling, the sampler repeatedly samples every possible boundary position, conditioned on the current state of the rest of the corpus. It stochastically decides whether the corresponding local area consists of a single word w_1 or two words $w_2 w_3$ ($w_1 = w_2.w_3$). The conditional probabilities can be derived from (1).

Collapsed sampling is known for slow convergence. This property is especially problematic in our settings where the initial segmentation is given by a morphological analyzer. Since the analyzer deterministically segments text using pre-defined parameters, the resultant segmentation is fairly consistent. Segmentation errors involving unknown words also occur in a regular way. Intuitively, we start with a local optimum although it is not too distant from the global optimum. The collapsed Gibbs sampler is easily entrapped by this local optimum. For this reason, the initial segmentation is usually chosen at random (Goldwater et al., 2009). Sentence-based block sampling is also susceptible to consistent initialization (Liang et al., 2010).

5.2 Type-based Sampling

To achieve fast convergence, we adopt a block sampling algorithm named type-based sampling (Liang et al., 2010). For the unigram model, a type-based sampler jointly samples multiple positions that share the same *type*. Two positions have the same type if the corresponding areas are both of the form w_1 or $w_2 w_3$. Type-based sampling takes advantage of the exchangeability of multiple positions with the same type. Given n positions with the same type, the sampler first samples the number of new boundaries m' ($0 \leq m' \leq n$), and then uniformly arranges m' boundaries out of n positions.

Type-based sampling has the ability to jump from a local optimum (e.g. consistently segmented) to another stable state (consistently unsegmented). While Liang et al. (2010) used random initialization, we take particular note of the possibility of efficiently correcting the consistent segmentation by the analyzer.

Type-based sampling is, however, not applicable to the bigram model for two reasons. The first problem is sparsity. For the bigram model, we need to consider adjacent words, w_l on the left and w_r on the right. This means that each type consists of three or four words, $w_l w_1 w_r$ or $w_l w_2 w_3 w_r$. Consequently, few positions share the same type and we fail to change closely-related areas $w_{l'} w_1 w_{r'}$ and $w_{l'} w_2 w_3 w_{r'}$, making inference inefficient.

The second and more fundamental problem arises from the hierarchical settings. Since the bigram model depends on latent table assignments, the joint distribution of multiple positions is no longer a closed-form function of counts.

Strictly speaking, we need to update the model counts even when sampling one position because the observation of the bigram $\langle w_l w_1 \rangle$, for example, may affect the probability $P_2(w_2 | \mathbf{h}_-, \langle w_l w_1 \rangle)$. Goldwater et al. (2009) approximate the probability by not updating the model counts in collapsed Gibbs

sampling (i.e. $P_2(w_2|\mathbf{h}_-, \langle w_l w_1 \rangle) \approx P_2(w_2|\mathbf{h}_-)$). They rely on the assumption that repeated bigrams are rare. Obviously this does not hold true for type-based sampling. Hence for type-based sampling, we have to update the model counts whenever we observe a new word.

One way to obtain the joint probability is to explicitly simulate the updates of histograms and other model counts. This is very cumbersome as we need to simulate $n + 1$ ways of model updates.

5.3 Hybrid Type-based Sampling

To address these problems, we propose a hybrid sampler which incorporates the Metropolis-Hastings algorithm into blocked Gibbs sampling. Metropolis-Hastings is another technique for sampling from a Markov chain. It first draws a proposed next state h' based on the current state h according to some proposal distribution $Q(h'; h)$. Then it accepts the proposal with the probability of

$$\min\left\{\frac{P(h')Q(h;h')}{P(h)Q(h';h)}, 1\right\}. \quad (4)$$

If the proposal is not accepted, the current state is used as the next state. Metropolis-Hastings is useful when it is difficult to directly sample from P.

We use the Metropolis-Hastings algorithm within Gibbs sampling. Instead of calculating the $n + 1$ probabilities of the number of boundaries, we only compare the current state with a proposed boundary arrangement. Also, the set of positions sampled jointly is chosen at unigram-level type instead of bigram-level type. The positions are no longer exchangeable. Therefore we calculate the conditional probability of one specific boundary arrangement.

When $n = 1$, the only choice is to flip the current state (i.e. $(m, m') \in \{(0, 1), (1, 0)\}$). This reduces to simple collapsed sampling. Otherwise we draw a proposed state in two steps. Given the n positions and the number of current boundaries m, we first draw the number of proposed boundaries m' from a probability distribution $f_n(m'; m)$. We then randomly arrange m' boundaries. The probability mass is uniformly divided by ${}_nC_{m'}$ arrangements. One exception is the case when $m \notin \{0, n\}$ and $m' = m$. In this case we perform permutation to obtain $h' \neq h$. To sum up, the proposal distribution

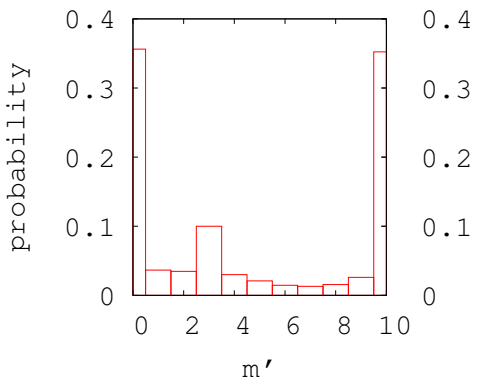

Figure 1: Probability of # of boundaries $f_{10}(m'; 3)$.

is defined as follows:

$$Q(h'; h) = \frac{f_n(m'; m)}{{}_nC_{m'} - I_n(m, m')}, \quad (5)$$

where $I_n(m, m')$ is 1 if $m \notin \{0, n\}$ and $m' = m$; otherwise 0.

We construct $f_n(m'; m)$ by discretizing a beta distribution ($\alpha = \beta < 1$) and a normal distribution with mean m, as shown in Figure 1. The former favors extreme values while the latter prefers smaller moves.

The sampling of each type is done in the following steps.

1. Collect n positions that share a unigram-level type.
2. Propose a new boundary arrangement. In what follows, we only focus on flipped boundaries because the rest does not change the likelihood ratio of the current and proposed states.
3. Calculate the current conditional probability. This can be done by repeatedly applying (2) while removing words one-by-one and updating the model counts accordingly.
4. Calculate the proposed conditional probability while adding words one-by-one.
5. Decide whether to accept the proposal according to (4). If the proposal is accepted, we finalize the arrangement; otherwise we revert to the current state.

We implement skip approximation (Liang et al., 2010) and sample each type once per iteration. This is motivated by the observation that although the

joint sampling of a large number of positions is computationally expensive, the proposal is accepted very infrequently.

5.4 Additional Constraints

Partial annotations (Tsuboi et al., 2008; Neubig and Mori, 2010) can be used for inference. If we know in advance that a certain position is a boundary or non-boundary, we simply keep it unaltered. As partially-annotated text, we can use markup. Suppose that the original text is written with wiki markup as follows:

*JR[[宇野線]][[常山駅]]
[gloss] JR Ube Line Tsuneyama Station

It is clear that the position between "線" (line) and "常" (*tsune*) is a boundary.

Similarly, we can impose our trivial rules of segmentation on the model. For example, we can keep punctuation markers (Li and Sun, 2009) separate from others.

6 Experiments

6.1 Settings

Data Set We evaluated our approach on Japanese Wikipedia. For each entry of Wikipedia, we regarded the title as a noun phrase and used both the title and main text for segmentation. We separately applied our segmentation procedure to each entry.

We constructed the data set as follows. We extracted each entry from an XML dump of Japanese Wikipedia.[4] We normalized the title by dropping trailing parentheses that disambiguate entries with similar names (e.g. "赤城 (空母)" for Akagi (aircraft carrier)). We extracted the main text from wikitext and used wiki markup as boundary markers. We applied both the title and main text to the morphological analyzer JUMAN[5] to get an initial segmentation. If the resultant segmentation conflicted with markup information, we overrode the former. The initial segmentation was also used as the baseline.

We only used entries that satisfied all of the following conditions.

1. The (normalized) title is longer than one character and contains hiragana, katakana and/or kanji.
2. The main text is longer than 1,000 characters.
3. The title appears at least 5 times in the main text.

The first condition ensures that there are segmentation ambiguities. The second and third conditions exclude entries unsuitable for statistical methods. 14% of the entries satisfied these conditions.

We randomly selected 500 entries and manually segmented their titles for evaluation. The 2-person inter-annotator Kappa score was 0.95.

As an annotated corpus, we used Kyoto Text Corpus.[6] It contained 1,675,188 characters.

Models We compared the unigram and bigram models. As for inference procedures, we used collapsed Gibbs sampling (**CL**) for both models, type-based sampling (**TB**) for the unigram model and hybrid type-based sampling (**HTB**) for the bigram model.

We tested two mixing methods of the annotated corpus, direct mixing (**DM**) and back-off mixing (**BM**).

To investigate the effect of initialization, we also tried randomly segmented text as the initial state (**RAND**). For random initialization, we placed a boundary with probability 0.5 on each position unless it was a fixed boundary.

The unigram model has one Dirichlet process concentration hyperparameter α_0 and the bigram model has α_0 and α_1. For each model, we experimented with the following values.
α_0: 0.1, 0.5, 1 5 10, 50, 100, 500, 1,000 and 5,000
α_1: 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100 and 500
For comparison, we also performed hyperparameter sampling. Following Escobar and West (1995), we set a gamma prior and introduced auxiliary variables to infer concentration parameters from data. For back-off mixing, we used the linear interpolation parameter $\lambda_{\mathrm{IP}} = 0.5$. The zerogram model was trained on the annotated corpus.

In each run, we performed 10 burn-in iterations. We then performed another 10 iterations to collect samples.

[4] http://download.wikimedia.org/jawiki/
[5] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[6] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?Kyoto%20University%20Text%20Corpus

Table 1: Results of segmentation of entry titles (F-score (precision/recall)).

model	best		median		inferred	
unigram + CL	81.35	(77.78/85.27)**	80.09	(75.80/84.89)	80.86	(76.81/85.36)
unigram + TB	55.87	(66.71/48.06)	51.04	(62.64/43.06)	42.63	(54.91/34.84)
bigram + CL	80.65	(76.73/84.99)	79.96	(75.50/84.99)	80.54	(76.84/84.61)
bigram + HTB	83.23	(85.25/81.30)**	74.52	(71.33/78.00)	34.52	(46.69/27.38)
unigram + CL + DM	85.29	(83.14/**87.54**)**	**81.62**	(**77.93**/**85.70**)**	80.91	(**82.87**/79.04)
unigram + TB + DM	35.26	(47.74/29.95)	33.81	(46.20/26.66)	31.90	(44.30/24.93)
bigram + CL + DM	80.37	(76.01/85.27)	79.88	(75.42/84.89)	73.77	(78.49/69.59)
bigram + HTB + DM	69.66	(67.68/71.77)	67.39	(64.35/70.73)	31.54	(43.79/24.64)
unigram + CL + BM	81.28	(77.48/85.46)	80.23	(76.06/84.89)	81.42	(77.75/**85.46**)
unigram + TB + BM	57.22	(68.01/49.39)	52.98	(64.50/44.95)	42.43	(54.69/34.66)
bigram + CL + BM	81.33	(77.34/85.74)	80.07	(75.69/84.99)	**81.46**	(77.82/**85.46**)**
bigram + HTB + BM	**86.32**	(**85.67**/86.97)**	76.35	(71.89/81.40)	40.81	(53.35/33.05)
unigram + TB + RAND	56.01	(66.93/48.16)	50.89	(62.21/43.06)	42.68	(54.81/34.94)
bigram + HTB + RAND	79.68	(80.13/79.23)	68.16	(63.64/73.37)	34.99	(47.05/27.86)
unigram + TB + BM + RAND	57.44	(67.91/49.76)	50.86	(61.92/43.15)	42.31	(54.55/34.56)
bigram + HTB + BM + RAND	84.03	(83.10/84.99)	70.46	(65.25/76.58)	40.16	(52.60/32.48)
baseline (JUMAN)	80.09	(75.80/84.89)				

** Statistically significant improvement with $p < 0.01$.

Evaluation Metrics We evaluated the segmentation accuracy of 500 entry titles. Specifically we evaluated the performance of a model with precision, recall and the F-score, all of which were based on tokens. We report the score of the most frequent segmentation among 10 samples.

Following Lee et al. (2010), we report the best and median settings of hyperparameters based on the F-score, in addition to inferred values.

In order to evaluate the degree of difference between a pair of segmentations, we employed character-based evaluation. Following Kudo et al. (2004), we converted a word sequence into character-based BI labels and examined labeling disagreements. McNemar's test of significance was based on this metric.

6.2 Results

Table 1 shows segmentation accuracy of various models. One would notice that the baseline score is much lower than the score previously reported regarding newspaper articles (Kudo et al., 2004). It is because unlike newspaper articles, the titles of Wikipedia entries contain an unusually high proportion of unknown words. As suggested by relatively low precision, unknown words tend to be over-segmented by the morphological analyzer.

In the best hyperparameter settings, the back-off mixing bigram model with hybrid type-based sampling (bigram + HTB + BM) significantly outperformed the baseline and achieved the best F-score. It did not performed well in the median setting as it was sensitive to the value of α_1. Hyperparameter estimation led to catastrophic decreases in bigram models as it made the hyperparameters much larger than those in the best settings.

Collapsed sampling (+CL) returned scores comparable to that of the baseline. It is simply because it did not change the initial segmentation a lot. In contrast, type-based sampling (+TB) brought large moves to the unigram model and significantly hurt accuracy. As suggested by relatively low recall, the unigram model prefers under-segmentation.

When combined with (hybrid) type-based sampling (+TB/+HTB), back-off mixing (+BM) increased accuracy from the corresponding non-mixing models. By contrast, direct mixing (+DM) drastically decreased accuracy from the non-mixing models. We can confirm that when the main text is orders of magnitude smaller than the annotated text, the role of constituent words in the main text is underestimated. To our surprise, collapsed sampling with mixing models (+CL, +DM/+BM) outperformed the baseline. However, the scores of type-based sampling (+TB) suggest that with much more iterations, the models would converge to undesired states.

The unigram model with random initialization was indifferent from that with default initialization. By contrast, the performance of the bigram model slightly degenerated with random initialization.

6.3 Convergence

Figure 2 shows how segmentations differed from the initial state in the course of inference.[7] A *diff* is defined as the number of character-based disagreements between the baseline segmentation and a model output. Hyperparameters used were those of the best model with (hybrid) type-based sampling.

We can see that collapsed sampling was almost unable to escape the initial state. With type-based sampling (+TB), the unigram model went further than the bigram model, but to an undesired direction. The bigram model with hybrid type-based sampling (bigram + HTB) converged in few iterations. Although the model with random initialization (+RAND) converged to a nearby point, the initial segmentation by the morphological analyzer realized a bit faster convergence and better accuracy.

Figure 2 shows how acceptance rates changed during inference. For comparison, a sample by a type-based Gibbs sampler was treated as "accepted" if the number of new boundaries was different from that of the current boundaries (i.e. $m' \neq m$). The acceptance rates were low and samplers seemingly stayed around modes.

6.4 Approximation

Up to this point, we consider every possible boundary position. However, this seems wasteful, given that a large portion of text has only marginal influence on the segmentation of the noun phrase in question. For this reason, we implemented approximation named matching skip. We sampled a boundary only if the corresponding local area contained a substring of the noun phrase in question.

Table 2 shows the result of approximation. Hyperparameters used were those of the best models with full sampling. Matching skip steadily worsened performance although not to a large extent. Meanwhile it drastically reduced the number of sampled positions. The median skip rate was 90.87%, with a standard deviation of 8.5.

[7] For a fair comparison, we might need to report changes over time instead of iterations. However, the difference of convergence speed is obvious in the iteration-based comparison although (hybrid) type-based sampling takes several times longer than collapsed sampling in the current naïve implementation.

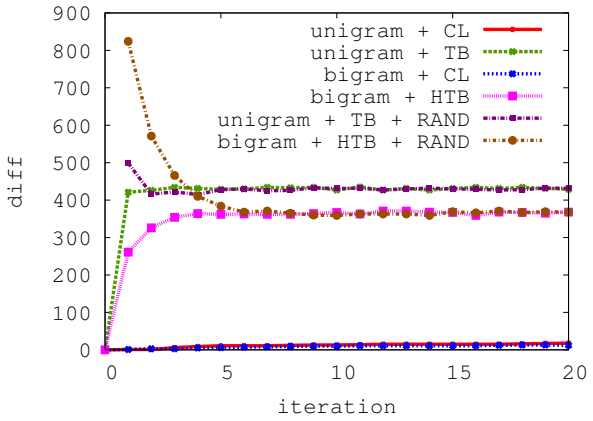

Figure 2: Diffs in the course of iteration. All models were with back-off mixing (+BM).

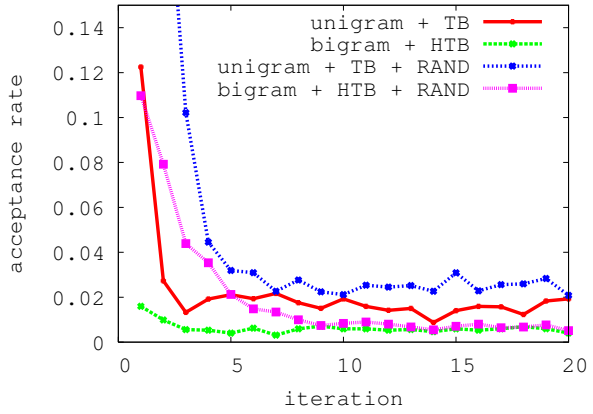

Figure 3: Acceptance rates for a noun phrase in the course of iteration. All models were with back-off mixing (+BM).

6.5 Discussion

Figure 4 shows some segmentations corrected by the back-off mixing bigram model with hybrid type-based sampling. "市比野" (*ichihino*) is a rare place name but can be identified by the model because it is frequently used in the article. "こなみるく" (*konamiruku* in hiragana) seems a pun on "粉ミルク" (*kona miruku*, "powdered milk") and "コナミ" (*konami* in katakana, a company). We consider it as a single word because we cannot reconstruct the etymology solely based on the main text. Note the different scripts. In Japanese, people often change the script to derive a proper noun from a common noun, which a naïve analyzer fails to recognize. It is

Table 2: Effect of matching skip (F-score (precision/recall)).

model	full		matching skip	
bigram + HTB	83.23	(85.25/81.30)**	82.86	(84.27/81.49)
bigram + HTB + BM	86.32	(85.67/86.97)**	83.87	(82.60/85.17)**
bigram + HTB + RAND	79.68	(80.13/79.23)	78.81	(78.64/75.07)
bigram + HTB + BM + RAND	84.03	(83.10/84.99)	81.08	(80.22/81.96)
baseline (JUMAN)	80.09	(75.80/84.89)		

** Statistically significant improvement with $p < 0.01$.

樋 + 脇 + 町 + 市 + 比 + 野 ⇒ 樋脇 + 町 + 市比野
hiwaki chou ichihino
(Ichihino, Hiwaki Town, an address)

り + そな + カード ⇒ りそな + カード
risona kaRdo
(Risona Card, a company)

ちり + とて + ちん ⇒ ちりとてちん
chiritotechiN (name of a play)

こな + みる + く ⇒ こなみるく
konamiruku
(a shop affiliated with Konami Corporation)

はい + じぃ ⇒ はいじぃ
haiziI (stage name of a comedian)

ちん + すこう ⇒ ちんすこう
chiNsukou (a traditional sweet)

コントラアルトクラリネット ⇒ コントラ + アルト + クラリネット
koNtora aruto kurarineQto (Contra-alto clarinet)

Figure 4: Examples of improved segmentations.

very important to identify hiragana words correctly. As hiragana is mainly used to write function words and other basic words, segmentation errors concerning hiragana often bring disastrous effects on applications of morphological analysis. For example, the analyzer over-segments "ちりとてちん" (*chiritotechiN*) into three shorter words among which the second word "とて" (*tote*) is a particle, and this sequence of words is transformed into a terrible parse tree.

Most improvements come from correction of over-segmentation because the initial segmentation by the analyzer shows a tendency of over-segmentation. An example of corrected under-segmentation is "contra-alto clarinet." The presence of "clarinet," "alto" and "contrabass" and others in the main text allowed the model to identify the constituents. On the other hand, the segmentation failed when our assumption about constituents does not hold. For example, the person name "菊池俊吉" (*kikuchi shuNkichi*) is two words but was erroneously combined into a single word by the model because unfortunately he was always referred to by the full name.

7 Conclusions

In this paper, we proposed a new task of Japanese noun phrase segmentation. We adopted non-parametric Bayesian language models and proposed hybrid type-based sampling that can efficiently correct segmentation given by the morphological analyzer. Although supervised segmentation is very competitive, we showed that it can be supplemented with our unsupervised approach.

We applied the proposed method to encyclopedic text to segment noun phrases in it. The proposed method can be applied to other tasks. For example, in unknown word acquisition (Murawaki and Kurohashi, 2008), noun phrases are often acquired from text as single words. We can now segment them into words in a more sophisticated way.

In the future we will assign a POS tag to each word in order to use segmented noun phrases in morphological analysis. We assume that the meaning of constituents in a noun phrase rarely depends on outer context. So it would be helpful to augment them with rich semantic information in advance instead of disambiguating their meaning every time we analyze given text.

Acknowledgments

This work was partly supported by JST CREST.

References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech

tagger. In *Proc. of COLING 2000*, pages 21–27.

Masayuki Asahara and Yuji Matsumoto. 2004. Japanese unknown word identification by character-based chunking. In *Proc. COLING 2004*, pages 459–465.

Phil Blunsom, Trevor Cohn, Sharon Goldwater, and Mark Johnson. 2009. A note on the implementation of hierarchical Dirichlet processes. In *Proc. of ACL-IJCNLP 2009: Short Papers*, pages 337–340.

Michael D. Escobar and Mike West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Thomas S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Interpolating between types and tokens by estimating power-law generators. In *NIPS 18*, pages 459–466.

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.

Mark Johnson and Katherine Demuth. 2010. Unsupervised phonemic Chinese word segmentation using adaptor grammars. In *Proc. of COLING 2010*, pages 528–536.

Mark Johnson. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proc. of ACL 2008*, pages 398–406.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proc. of ACL 2008*, pages 407–415, June.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proc. of EMNLP 2004*, pages 230–237.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proc. of The International Workshop on Sharable Natural Language Resources*, pages 22–38.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2010. Simple type-level unsupervised POS tagging. In *Proc. of EMNLP 2010*, pages 853–861.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for Chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

Percy Liang, Michael I. Jordan, and Dan Klein. 2010. Type-based MCMC. In *Proc. of NAACL 2010*, pages 573–581.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proc. of the 4th SIGHAN Workshop*, pages 161–164.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proc. of ACL-IJCNLP 2009*, pages 100–108.

Yugo Murawaki and Sadao Kurohashi. 2008. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proc. of EMNLP 2008*, pages 429–437.

Masaaki Nagata. 1996. Automatic extraction of new words from Japanese texts using generalized forward-backward search. In *Proc. of EMNLP 1996*, pages 48–59.

Tetsuji Nakagawa and Yuji Matsumoto. 2006. Guessing parts-of-speech of unknown words using global information. In *Proc. of COLING-ACL 2006*, pages 705–712.

Graham Neubig and Shinsuke Mori. 2010. Word-based partial annotation for efficient corpus construction. In *Proc. of LREC 2010*.

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for machine translation. In *Proc. of COLING 2010*, pages 815–823.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proc. of COLING '04*, pages 562–568.

Yee Whye Teh. 2006. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proc. of COLING 2008*, pages 897–904.

Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. 2001. The unknown word problem: a morphological analysis of Japanese using maximum entropy aided by a dictionary. In *Proc. of EMNLP 2001*, pages 91–99.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proc. of COLING 2008*, pages 1017–1024.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

Toshio Yokoi. 1995. The EDR electronic dictionary. *Communications of the ACM*, 38(11):42–44.