Nonparametric Model for Inupiaq Morphology Tokenization

ThuyLinh Nguyen[1] *Stephan Vogel*[2]
(1) Carnegie Mellon University.
(2) Qatar Foundation, Qatar.
thuylinh@cs.cmu.edu, svogel@qf.org.qa

Abstract

We present how to use English translation for unsupervised word segmentation of low resource languages. The inference uses a dynamic programming algorithm for efficient blocked Gibbs sampling. We apply the model to Inupiaq morphology analysis and get better results than monolingual model as well as Morfessor output.

Keywords: Nonparametric model, Gibbs sampling, morphology tokenization.

Proceedings of COLING 2012: Demonstration Papers, pages 337–344,
COLING 2012, Mumbai, December 2012.

1 Introduction

The tasks of morphological analysis or word segmentation have relied on word-formation rules or on available pretokenized corpora such as a Chinese word list, the Arabic or Korean treebanks, or the Czech dependancy treebank. However, these resources are expensive to obtain and for small or endangered languages, these resources are even not available. In recent years, unsupervised methods have been developed to infer the segmentation from unlabeled data such using minimal description length (Creutz and Lagus, 2007), log-linear model (Poon et al., 2009), nonparametric Bayesian model (Goldwater et al., 2009).

1.	Aŋun maqpiġaaliuġaa Aiviq .	*the man is writing the book for Aiviq .*
	Aŋun(*man*) maqpiġaa(*book*) liuġ(*writing*) aa Aiviq(*Aiviq*) .	
2.	Aŋun maqpiġaaliuŋitkaa Aiviq .	*the man is not writing the book for Aiviq .*
	Aŋun(*man*) maqpiġaa(*book*) liu(*writing*) ŋit(*not*) kaa Aiviq(*Aiviq*) .	

Table 1: Two examples of Inupiaq text, their English translations and their morphology tokenizations and English alignments.

A different promising approach is using information from a second language to learn the morphology analysis of the low resource language. Look at the example of Inupiaq[1] and its English translation in Table 1. Without knowing the language, let alone its morphology, we can conjecture that Inupiaq morpheme equivalent to English's "not" must be a substring of "maqpiġaaliuŋitkaa" and be overlaping with"ŋitk". This derivation is not possible without English translation.

This paper presents a nonparametric model and blocked Gibbs sampling inference to automatically derive morphology analysis of a text through its English translation. [2](Snyder and Barzilay, 2008) also applied a Bayesian model with Dirichlet Process priors to multilingual word segmentation task. Their prior distribution of souce word and target word alignment is defined based on the phonetic matching between two words. The bilingual word segmentation therefore benefits only when the source language and the target language belongs to the same family but does not achieve the same benefit when two languages are unrelated, such as using English translation to segment Arabic or Hebrew. We model the base distribution of target-source word alignment depends on the cooccurance of the two words in parallel corpora, independent of the language pair, the experiment results show the benefit of using English translation for Inupiaq morphology analysis.

Our model inherits (Nguyen et al., 2010)'s model of joint distribution of tokenized source text and its alignment to the target language. The alignment consists of at most one-to-one mappings between the source and target words with null alignments possible on both sides. To get samples of the posterior distribution, we use blocked Gibbs sampling algorithm and sample the whole source sentence and its alignment at the same time. The inference uses the dynamic programming method to avoid explicitly considering all possible segmentations of the source sentence and their alignments with the target sentence. Our inference technique is an extension of the forward sampling-backward filtering presented by (Mochihashi et al., 2009) for monolingual word segmentation. Dynamic programming algorithm has been also employed to sample PCFG parse trees (Johnson et al., 2007) and grammar-based word segmentation (Johnson and Goldwater, 2009).

[1]Inupiaq is a language spoken in Northern Alaska

[2] We use the term *bilingual word segmentation* for the problem we are working on to differentiate with *monolingual word segmentation* problem of learning to segment the text without translation reference.

In the next section we will discuss the model. Section 3 will describe the inference in detail. Experiments and results will be presented in section 4.

2 Model

A source sentence $\bar{\mathbf{s}}$ is a sequence of $|\bar{\mathbf{s}}|$ characters $\left(c_1, c_2 \dots c_{|\bar{\mathbf{s}}|}\right)$. A segmentation of $\bar{\mathbf{s}}$ is a sentence $\mathbf{s}$ of $|\mathbf{s}|$ words $s_i{:}\left(s_1, \dots, s_{|\mathbf{s}|}\right)$. We use the notation with bar on top $\bar{\mathbf{s}}$ to denote a sequence of characters to distinguish with a sequence of segmented words $\mathbf{s}$. In sentence $\mathbf{s}$, the set of aligned words is $\mathbf{s}_{\text{al}}$ the set of nonaligned words is $\mathbf{s}_{\text{nal}}$, each word in $\mathbf{s}_{\text{al}}$ aligns to a word in the set of aligned target words $\mathbf{t}_{\text{al}}$, the set of nonaligned target words is $\mathbf{t}_{\text{nal}}$.

In the first example Inupiaq segmentation in Table 1, the segmented $\mathbf{s}$ = "Aŋun maqpiġaa liuġ aa Aiviq ." the set $\mathbf{s}_{\text{al}} = \{$Aŋun, maqpiġaa liuġ, Aviq$\}$, $\mathbf{s}_{\text{nal}} = \{aa\}$, the mapping of $\mathbf{s}_{\text{al}}$ and $\mathbf{t}_{\text{al}}$ is { Aŋun(*man*), maqpiġaa(*book*), liuġ(*writing*), Aiviq(*Aiviq*)}, the set of unaligned target is $\mathbf{t}_{\text{nal}} = \{$the, the, for $\}$.

The generative process of a sentence $\mathbf{s}$ and its alignment $\mathbf{a}$ consists of two steps. In the first step, $\mathbf{s}$ and the alignments of each word in $\mathbf{s}$ are generated, each source word is either null aligned or aligned to a target word. In the second step, the model generates null aligned target words. More specifically, the two steps generative process of $(\mathbf{s}, \mathbf{a})$ as follows:

1. Repeatedly generate word s_i and its alignment:
   - Generate the word s_i with probability $\mathsf{p}\left(s_i\right)$.
   - Mark s_i as not aligned with probability $\mathsf{p}\left(\mathsf{null} \mid s_i\right)$ or aligned with probability $\left(1 - \mathsf{p}\left(\mathsf{null} \mid s_i\right)\right)$.
     - If s_i is aligned, $s \in \mathbf{s}_{\text{al}}$, generate its aligned target word $t_{a_i} \in \mathbf{t}_{\text{al}}$ with probability $\mathsf{p}\left(t_{a_i} | s_i\right)$.
2. Generate the set of non-aligned target words $\mathbf{t}_{\text{nal}}$ given aligned target words $\mathbf{t}_{\text{al}}$ with probability $\mathsf{p}\left(\mathbf{t}_{\text{nal}} | \mathbf{t}_{\text{al}}\right)$.

We model the source word distibution $\mathsf{p}(s)$, null aligned source word $\mathsf{p}(\mathsf{null} \mid s)$ and null aligned target words $\mathsf{p}\left(\mathbf{t}_{\text{nal}} | \mathbf{t}_{\text{al}}\right)$ similar to the same model described in (Nguyen et al., 2010). The main difference is in how we define the base distribution of a target word given a source word $\mathsf{p}(t | s)$.

2.1 Source-Target Alignment Model

The probability $\mathsf{p}(t|s)$ that a target word t aligns to a source word s is drawn from a Pitman-yor process $t \mid s \sim \text{PY}\left(d, \alpha, \mathsf{p}_0(t|s)\right)$ here d and α are the input parameters, and $\mathsf{p}_0(t|s)$ is the base distribution.

In word alignment model, highly co-occuring word pairs are likely to be the translation of each other, this is the intuition behind all the statistical word alignment model. However, the standard IBM word alignment models are not applicable to define conditional distribution of a pair of sequence of characters as the source side and a target word. We propose source-target alignment base distribution that captures the cooccurence of a sequence of source characters $\bar{c} = \left(c_1, \dots c_{|\bar{c}|}\right)$ and a target word t as follows.

Given the sentence pair $(\mathbf{s}, \mathbf{t})$, let $\mathsf{count}(\bar{c}, t)_{(\mathbf{s},\mathbf{t})}$ be the number of times the pair $(\bar{c}, t)$ cooccur. In Table 1, the example 1 has $\mathsf{count}(\text{aa}, the)_{(\mathbf{s},\mathbf{t})} = 2$; $\mathsf{count}(\text{maqpiga}, book)_{(\mathbf{s},\mathbf{t})} = 1$.

$$\alpha\left(l,k,\mathbf{t_1},t\right)=\begin{cases}\mathsf{p}\left(\mathsf{null}\mid c_{l-k+1}^{l}\right)\mathsf{p}\left(c_{l-k+1}^{l}\right)\sum_{u=1}^{l-k}\sum_{\substack{t'\in\mathbf{t_1}\\ \text{or } t'=\mathsf{null}}}\alpha\left(l-k,u,\mathbf{t_1}\setminus t',t'\right) & \text{if } t=\mathsf{null}\\ \left(1-\mathsf{p}\left(\mathsf{null}\mid c_{l-k+1}^{l}\right)\right)\mathsf{p}\left(c_{l-k+1}^{l}\right)\mathsf{p}\left(t\mid c_{l-k+1}^{l}\right) & \text{if } t \text{ is not null}\\ \quad\sum_{u=1}^{l-k}\sum_{\substack{t'\in\mathbf{t_1}\setminus t\\ \text{or } t'=\mathsf{null}}}\alpha\left(l-k,u,\mathbf{t_1}\setminus t',t'\right) & \end{cases}$$

Figure 1: Forward function $\alpha\left(l,k,\mathbf{t_1},t\right)$.

Given the training parallel corpora $\mathcal{T}$ number of times $(\bar{c},t)$ cooccur in $\mathcal{T}$ is $\mathsf{count}(\bar{c},t)_{\mathcal{T}}=\sum_{(\mathbf{s},\mathbf{t})\in\mathcal{T}}\mathsf{count}(\bar{c},t)_{(\mathbf{s},\mathbf{t})}$.

Note that if "maqpiga" is the translation of "book" by cooccuring in sentence pairs, any substring of the word such as "m", "ap" also coocur with "*book*", but *book* is not their translation candidate. We therefore remove these counts from corpora coocurence count. Let $\bar{c'}$ be a substring of $\bar{c}$, if $\mathsf{count}(\bar{c'},t)_{\mathcal{T}}=\mathsf{count}(\bar{c},t)_{\mathcal{T}}$, $\mathsf{count}(\bar{c'},t)_{\mathcal{T}}>\theta_1$ and $\mathsf{length}\left(\bar{c'}\right)<\theta_2$, we remove $\mathsf{count}(\bar{c'},t)_{\mathcal{T}}$ from the count, $\mathsf{count}(\bar{c'},t)=0$. Here θ_1 and θ_2 are the input thresholds to avoid removal of low frequency words or long words. We define the base distribution for any sequence of character $\bar{c}$ and a target word t as: $\mathsf{p}_0(t\mid\bar{c})=\frac{\mathsf{count}(\bar{c},t)}{\sum_t\mathsf{count}(\bar{c},t)}$.

3 Inference

We use blocked Gibbs sampling algorithm to generate samples from the posterior distribution of the model. In our experiment, the variables are potential word boundaries and sentence alignments. We first initialize the training data with an initial segmentation and alignment. Then the sampler iteratively removes a sentence from the data and samples the segmentation and alignment of the sentence given the rest of training set. The new sentence is then added to the corpus. This process continues until the samples are mixed up and represent the posterior of interest.

We apply the dynamic programming Forward filtering - Backward sampling algorithm. The Forward filtering step calculates the marginalized probabilities of variables and the Backward sampling step uses these probabilities to sample the variables.

3.1 Forward Filtering

We define the forward function $\alpha\left(l,k,\mathbf{t_1},t\right)$ to iteratively calculate the marginalized probabilities in step 1 of generation process. That is the probability of the substring c_1^l with the final k characters being a word aligned to t and c_1^{l-k} aligns to the subset target words $\mathbf{t_1}$.

Figure 1 shows how to calculate $\alpha\left(l,k,\mathbf{t_1},t\right)$, the mathematic derivation of the function is similar to forward function of HMM model. We will give the detail in a technical report.

3.2 Backward Sampling

The forward filtering step calculates function $\alpha\left(l,k,\mathbf{t_1},t\right)$ in the order of increasing l. At the end of the source sentence, the forward filtering step calculates the forward variables $\alpha\left(|\mathbf{s}|,k,\mathbf{t_1},t\right)$ for any $k\leq|\mathbf{s}|$, $\mathbf{t_1}\subset\mathbf{t}$ and $t\in\mathbf{t}$, $t\notin\mathbf{t_1}$. That is the probability that the sentence ~~s~~ has the last

word with length k and the word aligns to t, the rest of the sentence aligns to the set $\mathbf{t_1}$. The set of corresponding aligned target words is $\mathbf{t_1} \cup \{t\}$.

So the marginalized probability in step 1 of the generation process is

$$\mathrm{p}\left(\cancel{\mathbf{s}}, \mathbf{t}_{\mathrm{al}}\right) = \sum_{k} \sum_{\substack{t' \in \mathbf{t_1} \setminus t \\ \text{or } t' = \mathsf{null}}} \alpha\left(\cancel{|\mathbf{s}|}, k, \mathbf{t_1} \setminus t', t'\right)$$

We have the marginalized probability that any $\mathbf{t}_{\mathrm{nal}} \subset \mathbf{t}$ is the set of nonaligned target word: $\mathrm{p}\left(\cancel{\mathbf{s}}, \mathbf{t}_{\mathrm{al}}, \mathbf{t}_{\mathrm{nal}}\right) = \mathrm{p}\left(\cancel{\mathbf{s}}, \mathbf{t}_{\mathrm{al}}\right) \times \mathrm{p}\left(\mathbf{t}_{\mathrm{nal}} \,|\, \mathbf{t}_{\mathrm{al}}\right)$.

In backward sampling, we first sample the set of nonaligned target words $\mathbf{t}_{\mathrm{nal}}$ proportional to the marginalized probability $\mathrm{p}\left(\cancel{\mathbf{s}}, \mathbf{t}_{\mathrm{al}}, \mathbf{t}_{\mathrm{nal}}\right)$.

Once we got the set of aligned target words $\mathbf{t}_{\mathrm{nal}}$ for string $\mathbf{s}$, we sample the length k of the last word of the sentence according to its marginalized probability:

$$\sum_{\substack{t \in \mathbf{t}_{\mathrm{al}} \\ \text{or } t = \mathsf{null}}} \alpha\left(\cancel{|\mathbf{s}|}, k, \mathbf{t}_{\mathrm{al}} \setminus t, t\right).$$

then sample the alignment t of the last word $c_{\cancel{|\mathbf{s}|}-k+1}^{\cancel{|\mathbf{s}|}}$ proportional to $\alpha\left(\cancel{|\mathbf{s}|}, k, \mathbf{t}_{\mathrm{al}} \setminus t, t\right)$. The process continues backward until we reach the beginning of the sentence.

4 Empirical Results

The Inupiaq-English data is from an ellicit parallel corpora which consists of 2933 parallel sentences of average 4.34 words per Inupiaq sentence, 6 words per English sentences. On average, each Inupiaq word has 11.35 characters. Our task is to segment Inupiaq into morphemes with average 2.31 morphemes per word. The number of morphemes per word has high variance, Inupiaq usually has subject and oblique as separate words and the rest of the sentence as one long word. It is a rich morphology language with relatively free reordering. For example, both sentences "Maqpiaaliuġniaġaa Aiviq Paniattaam" and "Aiviq maqpiaaliuġniaġaa Paniattaam" have the same translation *"Paniattaaq will write a book for Aiviq"*.

4.1 Experiment Setup

We report the best result of experiments with different values hyperparameter p_0 (from $0.5 \ldots 0.9$ and 0.95, 0.99) which model the source word null alignment and the parameter represent word length distribution. All other hyperparameters are fixed before the experiment. The MCMC inference algorithm starts with an initial segmentation as full word form and initial monotone alignment. The inference samples 100 iterations through the whole training set. The inference adopts simulated annealing to speed convergence of the sampler and approximates the samples around the MAP solution. In the last 10 iterations, the inference uses a temperature parameter τ and starts cooling down from 1, $\frac{9}{10}$ to $\frac{1}{10}$. The last iteration represents the MAP solution and is the segmentation output. We leave the analysis of inference convergence to the future work. We report the segmentation accuracy on F-score of:[3] **BF**: word boundries score; **F**: word token score: both boundaries of a word must be correctly identified to be counted as correct; **LF**: instead of reporting the word-token accuracy as F, LF represents word-type score.

[3] the evaluation script is on http://homepages.inf.ed.ac.uk/sgwater/resources.html

One word in rich morphology language is often equivalent to several English words. But English has morphology in the language too, an English past tense verb or plural noun are equivalent to two morphemes. We also experiment our bilingual model with the target side is tokenized English. The tokenized English is the result pf manually convert English past tense verbs into "verb PAST", plural nouns into "lematize_noun PL", all other words are lemmatized. For example, the sentence *The men sang* is converted to *The man PL sing PAST*.

4.2 Empirical Results

	F	BF	LF
HDP-Monolingual	35.02	62.62	**30.51**
Morfessor	35.53	63.98	29.72
Bilingual	35.83	65.68	29.80
Bilingual-tokenized Eng	**39.15**	**66.71**	30.15

Table 2: Inupiaq morphology analysis results: comparing the baseline scores (upper row) and our scores (lower row).

Table 2 presents our segmentation results. The top rows are monolingual baselines, the bottom rows are our bilingual results using either English or tokenized English as the target language. The best score belongs to bilingual model using tokenized English for unsupervised segmentation.

For monolingual baselines, we use (Goldwater et al., 2009)'s monolingual nonparametric word segmentation. The HDP-Monolingual result in Table 2 is the best score from nearly 100 experiments of their model with different hyperparameters. Unlike (Snyder and Barzilay, 2008) report that bilingual word segmentation do not benefit when two languages are different. Our bilingual model still has better performance than the nonparametric monolingual word segmentation on F, BF and average scores. The informative source-target base distribution contribute to this better result.

We also use Morfessor (Creutz and Lagus, 2007), an unsupervised morphology analysis as another monolingual baseline. Morfessor is the state-of-the-art unsupervised segmentation for complex morphology languages such as Finish, Czech. The bilingual model using tokenized English significantly outperforms Morfessor result. Our experiment on small Inupiaq-English data is typical for low resource language scenario. The morphology analysis of the parallel data then can be used as supervised data for monolingual segmentation task without given English translation.

5 Conclusion

We have presented a bilingual nonparametric word segmentation model for low resource languages. The inference uses a dynamic programming algorithm for an efficient blocked Gibbs sampling. The experiment shows the benefit of using translation in word segmentation task.

References

Creutz, M. and Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):1–34.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition*, 112(1):21–54.

Johnson, M. and Goldwater, S. (2009). Improving Nonparameteric Bayesian Inference: Experiments on Unsupervised Word Segmentation with Adaptor Grammars. In *Proceedings of*

NAACL-HLT '09, pages 317–325, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johnson, M., Griffiths, T., and Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of NAACL-HLT '07*, pages 139–146, Rochester, New York. Association for Computational Linguistics.

Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proceedings of ACL:09*, pages 100–108, Suntec, Singapore. Association for Computational Linguistics.

Nguyen, T., Vogel, S., and Smith, N. A. (2010). Nonparametric Word Segmentation for Machine Translation. In *Proceedings of Coling-10*, pages 815–823, Beijing, China.

Poon, H., Cherry, C., and Toutanova, K. (2009). Unsupervised Morphological Segmentation with Log-Linear Models. In *Proceedings NAACL'09*, pages 209–217.

Snyder, B. and Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio. Association for Computational Linguistics.