# Empirical Study of Unsupervised Chinese Word Segmentation Methods for SMT on Large-scale Corpora

**Xiaolin Wang** **Masao Utiyama** **Andrew Finch** **Eiichiro Sumita**
National Institute of Information and Communications Technology
{xiaolin.wang,mutiyama,andrew.finch,eiichiro.sumita}@nict.go.jp

## Abstract

Unsupervised word segmentation (UWS) can provide domain-adaptive segmentation for statistical machine translation (SMT) without annotated data, and bilingual UWS can even optimize segmentation for alignment. Monolingual UWS approaches of explicitly modeling the probabilities of words through Dirichlet process (DP) models or Pitman-Yor process (PYP) models have achieved high accuracy, but their bilingual counterparts have only been carried out on small corpora such as basic travel expression corpus (BTEC) due to the computational complexity. This paper proposes an efficient unified PYP-based monolingual and bilingual UWS method. Experimental results show that the proposed method is comparable to supervised segmenters on the in-domain NIST OpenMT corpus, and yields a 0.96 BLEU relative increase on NTCIR PatentMT corpus which is out-of-domain.

## 1 Introduction

Many languages, especially Asian languages such as Chinese, Japanese and Myanmar, have no explicit word boundaries, thus word segmentation (WS), that is, segmenting the continuous texts of these languages into isolated words, is a prerequisite for many natural language processing applications including SMT.

Though supervised-learning approaches which involve training segmenters on manually segmented corpora are widely used (Chang et al., 2008), yet the criteria for manually annotating words are arbitrary, and the available annotated corpora are limited in both quantity and genre variety. For example, in machine translation, there are various parallel corpora such as BTEC for tourism-related dialogues (Paul, 2008) and PatentMT in the patent domain (Goto et al., 2011)[1], but researchers working on Chinese-related tasks often use the Stanford Chinese segmenter (Tseng et al., 2005) which is trained on a small amount of annotated news text.

In contrast, UWS, spurred by the findings that infants are able to use statistical cues to determine word boundaries (Saffran et al., 1996), relies on statistical criteria instead of manually crafted standards. UWS learns from unsegmented raw text, which are available in large quantities, and thus it has the potential to provide more accurate and adaptive segmentation than supervised approaches with less development effort being required.

The approaches of explicitly modeling the probability of words(Brent, 1999; Venkataraman, 2001; Goldwater et al., 2006; Goldwater et al., 2009; Mochihashi et al., 2009) significantly outperformed a heuristic approach (Zhao and Kit, 2008) on the monolingual Chinese SIGHAN-MSR corpus (Emerson, 2005), which inspired the work of this paper.

However, bilingual approaches that model word probabilities suffer from computational complexity. Xu et al. (2008) proposed a bilingual method by adding alignment into the generative model, but was only able to test it on small-scale BTEC data. Nguyen et al. (2010) used the local best alignment to increase the speed of the Gibbs sampling in training but the impact on accuracy was not explored.

This paper is dedicated to bilingual UWS on large-scale corpora to support SMT. To this end, we model bilingual UWS under a similar framework with monolingual UWS in order to improve efficiency, and replace Gibbs sampling with expectation maximization (EM) in training.

We aware that variational bayes (VB) may be used for speeding up the training of DP-based

[1] http://ntcir.nii.ac.jp/PatentMT



*Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 752–758,
Baltimore, Maryland, USA, June 23-25 2014. 

or PYP-based bilingual UWS. However, VB requires formulating the $m$ expectations of $(m-1)$-dimensional marginal distributions, where $m$ is the number of hidden variables. For UWS, the hidden variables are indicators that identify substrings of sentences in the corpus as words. These variables are large in number and it is not clear how to apply VB to UWS, and as far the authors aware there is no previous work related to the application of VB to monolingual UWS. Therefore, we have not explored VB methods in this paper, but we do show that our method is superior to the existing methods.

The contributions of this paper include,

- state-of-the-art accuracy in monolingual UWS;
- the first bilingual UWS method practical for large corpora;
- improvement of BLEU scores compared to supervised Stanford Chinese word segmenter.

## 2 Methods

This section describes our unified monolingual and bilingual UWS scheme. Table 1 lists the main notation. The set $\mathcal{F}$ is chosen to represent an unsegmented foreign language sentence (a sequence of characters), because an unsegmented sentence can be seen as the set of all possible segmentations of the sentence denoted $F$, i.e. $F \in \mathcal{F}$.

| Notation | Meaning |
|---|---|
| $\mathcal{F}$ | an unsegmented foreign sentence |
| $\mathcal{F}_k^{k'}$ | unsegmented substring of the underlying string of $\mathcal{F}$ from $k$ to $k'$ |
| $F$ | a segmented foreign sentence |
| $f_j$ | the $j$-th foreign word |
| $\mathcal{M}$ | monolingual segmentation model |
| $P_{\mathcal{M}}(x)$ | probability of $x$ being a word according to $M$ |
| $E$ | a tokenized English sentence |
| $e_i$ | the $i$-th English word |
| ($\mathcal{F}$,$E$) | a bilingual sentence pair |
| $\mathcal{B}$ | bilingual segmentation model |
| $P_{\mathcal{B}}(x\|e_i)$ | probability of $x$ being a word according to $B$ given $e_i$ |

Table 1: Main Notation.

Monolingual and bilingual WS can be formulated as follows, respectively,

$$\hat{F}(\mathcal{F}) = \underset{F \in \mathcal{F}}{argmax}\, P(F|\mathcal{F}, \mathcal{M}), \quad (1)$$

$$\hat{F}(\mathcal{F}, E) = \underset{F \in \mathcal{F}}{argmax} \sum_a P(F, a|\mathcal{F}, E, \mathcal{B}), \quad (2)$$

where $a$ is an alignment between $F$ and $E$. The English sentence $E$ is used in the generation of a segmented sentence $F$.

UWS learns models by maximizing the likelihood of the unsegmented corpus, formulated as,

$$\hat{\mathcal{M}} = \underset{\mathcal{M}}{\text{argmax}} \prod_{\mathcal{F} \in \mathbb{F}} \Big( \sum_{F \in \mathcal{F}} P(F|\mathcal{M}) \Big), \quad (3)$$

$$\hat{\mathcal{B}} = \underset{\mathcal{B}}{\text{argmax}} \prod_{(\mathcal{F},E) \in \mathbb{B}} \Big( \sum_{F \in \mathcal{F}} \sum_a P(F, a|\mathcal{F}, E, \mathcal{B}) \Big). \quad (4)$$

Our method of learning $\mathcal{M}$ and $\mathcal{B}$ proceeds in a similar manner to the EM algorithm. The following two operations are performed iteratively for each sentence (pair).

- Exclude the previous expected counts of the current sentence (pair) from the model, and then derive the current sentence in all possible ways, calculating the new expected counts for the words (see Section 2.1), that is, we calculate the expected probabilities of the $\mathcal{F}_k^{k'}$ being words given the data excluding $\mathcal{F}$, i.e. $\mathbf{E}_{\mathbb{F}/\{\mathcal{F}\}}(P(\mathcal{F}_k^{k'}|\mathcal{F})) = P(\mathcal{F}_k^{k'}|\mathcal{F}, \mathcal{M})$ in a similar manner to the marginalization in the Gibbs sampling process which we are replacing;
- Update the respective model $\mathcal{M}$ or $\mathcal{B}$ according to these expectations (see Section2.2).

### 2.1 Expectation

#### 2.1.1 Monolingual Expectation

$P(\mathcal{F}_k^{k'}|\mathcal{F}, \mathcal{M})$ is the marginal probability of all the possible $F \in \mathcal{F}$ that contain $\mathcal{F}_k^{k'}$ as a word, which can be calculated efficiently through dynamic programming (the process is similar to the foreward-backward algorithm in training a hidden Markov model (HMM) (Rabiner, 1989)):

$$P_a(k) = \sum_{u=1}^{U} P_a(k-u) P_{\mathcal{M}}(\mathcal{F}_{k-u}^{k})$$

$$P_b(k') = \sum_{u=1}^{U} P_b(k'+u) P_{\mathcal{M}}(\mathcal{F}_{k'}^{k'+u})$$

$$P(\mathcal{F}_k^{k'}|\mathcal{F}, \mathcal{M}) = P_a(k) P_{\mathcal{M}}(\mathcal{F}_k^{k'}) P_b(k'), \quad (5)$$

where $U$ is the predefined maximum length of foreign language words, $P_a(k)$ and $P_b(k')$ are the forward and backward probabilities, respectively. This section uses a unigram model for description convenience, but the method can be extended to $n$-gram models.

### 2.1.2 Bilingual Expectation

$P(\mathcal{F}_k^{k'}|\mathcal{F}, E, \mathcal{B})$ is the marginal probability of all the possible $F \in \mathcal{F}$ that contain $\mathcal{F}_k^{k'}$ as a word and are aligned with $E$, formulated as:

$$\begin{aligned} P(\mathcal{F}_k^{k'}|\mathcal{F}, E, \mathcal{B}) &= \sum_{\substack{F\in\mathcal{F} \\ \mathcal{F}_k^{k'}\in F}} \sum_a P(F, a|E, \mathcal{B}) \\ &\approx \sum_{\substack{F\in\mathcal{F} \\ F_{j_k}=\mathcal{F}_k^{k'}}} \sum_a \prod_{j=1}^{J} P(a_j|j, I, J) P_{\mathcal{B}}(f_j|e_{a_j}) \\ &= \sum_{\substack{F\in\mathcal{F} \\ f_{j_k}=\mathcal{F}_k^{k'}}} \prod_{j=1}^{J} \sum_a P(a_j|j, I, J) P_{\mathcal{B}}(f_j|e_{a_j}), \end{aligned} \tag{6}$$

where $J$ and $I$ are the number of foreign and English words, respectively, and $a_j$ is the position of the English word that is aligned to $f_j$ in the alignment $a$. For the alignment we employ an approximation to IBM model 2 (Brown et al., 1993; Och and Ney, 2003) described below.

We define the conditional probability of $f_j$ given the corresponding English sentence $E$ and the model $\mathcal{B}$ as:

$$P_{\mathcal{B}}(f_j|E) = \sum_a P(a_j|j, I, J) P_{\mathcal{B}}(f_j|e_{a_j}) \tag{7}$$

Then, the previous dynamic programming method can be extended to the bilingual expectation

$$\begin{aligned} P_a(k|E) &= \sum_{u=1}^{U} P_a(k-u|E) P_{\mathcal{B}}(\mathcal{F}_{k-u}^{k}|E) \\ P_b(k'|E) &= \sum_{u=1}^{U} P_b(k'+u|E) P_{\mathcal{B}}(\mathcal{F}_{k'}^{k'+u}|E) \\ P(F_k^{k'}|\mathcal{F}, E, \mathcal{B}) &= P_a(k|E) P_{\mathcal{B}}(\mathcal{F}_k^{k'}|E) P_b(k'|E). \end{aligned} \tag{8}$$

Eq. 7 can be rewritten (as in IBM model 2):

$$P_{\mathcal{B}}(f_j|E) = \sum_{i=1}^{I} P^*(i|j, I, J) P_{\mathcal{B}}(f_j|e_i) \tag{9}$$

$$P^*(i|j, I, J) = \sum_{a:a_j=i} P(a_j|, j, I, J)$$

In order to maintain both speed and accuracy, the following window function is adopted

$$P^*(i|j, I, J) \approx P^*(i|k, I, K) = \begin{cases} e^{-|i-kI/K|}/\sigma & |i - kI/K| \leqslant \delta_b/2 \\ \lambda_\phi & e_i \text{ is empty word} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $K$ is the number of characters in $\mathcal{F}$, and the $k$-th character is the start of the word $f_j$, since $j$ and $J$ are unknown during the computation of dynamic programming. $\delta_b$ is the window size, $\lambda_\phi$ is the prior probability of an empty English word, and $\sigma$ ensures all the items sum to 1.

## 2.2 Maximization

Inspired by (Teh, 2006; Mochihashi et al., 2009; Neubig et al., 2010; Teh and Jordan, 2010), we employ a Pitman-Yor process model to build the segmentation model $\mathcal{M}$ or $\mathcal{B}$. The monolingual model $\mathcal{M}$ is

$$\begin{aligned} P_{\mathcal{M}}(f_j) &= \frac{\max\big(n(f_j) - d, 0\big) + (\theta + d \cdot n_{\mathcal{M}}) G_0(f_j)}{\sum_{f'_j} n(f'_j) + \theta} \\ n_{\mathcal{M}} &= \big|\{f_j \,|\, n(f_j) \geqslant d\}\big|, \end{aligned} \tag{11}$$

where $f_j$ is a foreign language word, and $n(f_j)$ is the observed counts of $f_j$, $\theta$ is named the strength parameter, $G_0(f_j)$ is named the base distribution of $f_j$, and $d$ is the discount.

The bilingual model is

$$\begin{aligned} P_{\mathcal{B}}(f_j|e_i) &= \frac{\max\big(n(f_j, e_i) - d, 0\big) + (\theta + d \cdot n_{e_i}) G_0(f_j|e_i)}{\sum_{f'_j} n(f'_j, e_i) + \theta} \\ n_{e_i} &= \big|\{x \,|\, n(x, e_i) \geqslant d\}\big|. \end{aligned} \tag{12}$$

In Eqs. 11 and 12,

$$n(f_j) = \sum_{\mathcal{F}\in\mathbb{F}} P(f_j|\mathcal{F}, \mathcal{M}) \tag{13}$$

$$n(f_j, e_i) = \sum_{(\mathcal{F},E)\in\mathbb{B}} P(f_j|\mathcal{F}, E, \mathcal{B}) \frac{P^*(i|j, I, J) P_{\mathcal{B}}(f_j|e_i)}{\sum_{i'=1}^{I} P^*(i'|j, I, J) P_{\mathcal{B}}(f_j|e_{i'})}. \tag{14}$$

## 3 Complexity Analysis

The computational complexity of our method is linear in the number of iterations, the size of the corpus, and the complexity of calculating the expectations on each sentence or sentence pair. In practical applications, the size of the corpus is fixed, and we found empirically that the number of iterations required by the proposed method for convergence is usually small (less than five iterations). We now look in more detail at the complexity of the expectation calculation in monolingual and bilingual models.

The monolingual expectation is calculated according to Eq. 5; the complexity is linear in the length of sentences and the square of the predefined maximum length of words. Thus its overall complexity is

$$O^{\text{unigram}}_{\text{monoling}} = O(N_{\text{i}}|\mathbb{F}|KU^2), \qquad (15)$$

where $N_{\text{i}}$ is the number of iterations, $K$ is the average number of characters per sentence, and $U$ is the predefined maximum length of words.

For the monolingual bigram model, the number of states in the HMM is $U$ times more than that of the monolingual unigram model, as the states at specific position of $F$ are not only related to the length of the current word, but also related to the length of the word before it. Thus its complexity is $U^2$ times the unigram model's complexity:

$$O^{\text{bigram}}_{\text{monoling}} = O(N_{\text{i}}|\mathbb{F}|KU^4). \qquad (16)$$

The bilingual expectation is given by Eq. 8, whose complexity is the same as the monolingual case. However, the complexity of calculating the transition probability, in Eqs. 9 and 10, is $O(\delta_b)$. Thus its overall complexity is:

$$O^{\text{unigram}}_{\text{biling}} = O(N_{\text{i}}|\mathbb{F}|KU^2\delta_b). \qquad (17)$$

## 4 Experiments

In this section, the proposed method is first validated on monolingual segmentation tasks, and then evaluated in the context of SMT to study whether the translation quality, measured by BLEU, can be improved.

### 4.1 Experimental Settings

#### 4.1.1 Experimental Corpora

Two monolingual corpora and two bilingual corpora are used (Table 2). CHILDES (MacWhinney and Snow, 1985) is the most common test corpus for UWS methods. The SIGHAN-MSR corpus (Emerson, 2005) consists of manually segmented simplified Chinese news text, released in the SIGHAN bakeoff 2005 shared tasks.

| Corpus | Type | # Sentences | # Characters |
|---|---|---|---|
| CHILDES | Mono. | 9,790 | 95,809 |
| SIGHAN-MSR | Mono. | 90,903 | 4,234,824 |
| OpenMT06 | Biling. | 437,004 | 19,692,605 |
| PatentMT9 | Biling. | 1,004,000 | 63,130,757 |

Table 2: Experimental Corpora

The first bilingual corpus: OpenMT06 was used in the NIST open machine translation 2006 Evaluation [2]. We removed the United Nations corpus and the traditional Chinese data sets from the constraint training resources. The data sets of NIST Eval 2002 to 2005 were used as the development for MERT tuning (Och, 2003). This data set mainly consists of news text [3]. PatentMT9 is from the shared task of NTCIR-9 patent machine translation . The training set consists of 1 million parallel sentences extracted from patent documents, and the development set and test set both consist of 2000 sentences.

#### 4.1.2 Performance Measurement and Baseline Methods

For the monolingual tasks, the $F_1$ score against the gold annotation is adopted to measure the accuracy. The results reported in related papers are listed for comparison.

For the bilingual tasks, the publicly available system of Moses (Koehn et al., 2007) with default settings is employed to perform machine translation, and BLEU (Papineni et al., 2002) was used to evaluate the quality. Character-based segmentation, LDC segmenter and Stanford Chinese segmenters were used as the baseline methods.

#### 4.1.3 Parameter settings

The parameters are tuned on held-out data sets. The maximum length of foreign language words is set to 4. For the PYP model, the base distribution adopts the formula in (Chung and Gildea, 2009), and the strength parameter is set to $1.0$, and the discount is set to $1.0 \times 10^{-6}$.

For bilingual segmentation,the size of the alignment window is set to $6$; the probability $\lambda_\phi$ of foreign language words being generated by an empty

[2] `http://www.itl.nist.gov/iad/mig/ /tests/mt/2006/`

[3] It also contains a small number of web blogs

| Method | Accuracy | | Time | |
|---|---|---|---|---|
| | CHILD. | MSR | CHILD. | MSR |
| NPY(bigram)[a] | 0.750 | 0.802 | 17 m | – |
| NPY(trigram)[a] | 0.757 | **0.807** | – | – |
| HDP(bigram)[b] | 0.723 | – | 10 h | – |
| Fitness[c] | – | 0.667 | – | – |
| Prop.(unigram) | 0.729 | 0.804 | 3 s | 50 s |
| Prop.(bigram) | **0.774** | 0.806 | 15 s | 2530 s |

[a] by (Mochihashi et al.,2009);
[b] by (Goldwater et al.,2009);
[c] by (Zhao and Kit, 2008).

Table 3: Results on Monolingual Corpora.

English word, was set to 0.3.

The training was started from assuming that there was no previous segmentations on each sentence (pair), and the number of iterations was fixed. It was set to 3 for the monolingual unigram model, and 2 for the bilingual unigram model, which provided slightly higher BLEU scores on the development set than the other settings. The monolingual bigram model, however, was slower to converge, so we started it from the segmentations of the unigram model, and using 10 iterations.

### 4.2 Monolingual Segmentation Results

In monolingual segmentation, the proposed methods with both unigram and bigram models were tested. Experimental results show that they are competitive to state-of-the-art baselines in both accuracy and speed (Table 3). Note that the comparison of speed is only for reference because the times are obtained from their respective papers.

### 4.3 Bilingual Segmentation Results

Table 4 presents the BLEU scores for Moses using different segmentation methods. Each experiment was performed three times. The proposed method with monolingual bigram model performed poorly on the Chinese monolingual segmentation task; thus, it was not tested. We intended to test (Mochihashi et al., 2009), but found it impracticable on large-scale corpora.

The experimental results show that the proposed UWS methods are comparable to the Stanford segmenters on the OpenMT06 corpus, while achieves a 0.96 BLEU increase on the PatentMT9 corpus. This is because this corpus is out-of-domain for the supervised segmenters. The CTB and PKU Stanford segmenter were both trained on annotated news text, which was the major domain of OpenMT06.

| Method | BLEU | |
|---|---|---|
| | OpenMT06 | PatentMT9 |
| Character | $29.50 \pm 0.03$ | $28.36 \pm 0.09$ |
| LDC | $31.33 \pm 0.10$ | $30.22 \pm 0.14$ |
| Stanford(CTB) | $\mathbf{31.68 \pm 0.25}$ | $30.77 \pm 0.13$ |
| Stanford(PKU) | $31.54 \pm 0.13$ | $30.86 \pm 0.04$ |
| Prop.(mono.) | $31.47 \pm 0.18$ | $31.62 \pm 0.06$ |
| Prop.(biling.) | $31.61 \pm 0.14$ | $\mathbf{31.73 \pm 0.05}$ |

Table 4: Results on Bilingual Corpora.

| Method | Time | |
|---|---|---|
| | OpenMT06 | PatentMT9 |
| Prop.(mono.) | 28 m | 1 h 01 m |
| Prop.(biling.) | 2 h 25 m | 5 h 02 m |

Table 5: Time Costs on Bilingual Corpora.

Table 5 presents the run times of the proposed methods on the bilingual corpora. The program is single threaded and implemented in C++. The time cost of the bilingual models is about 5 times that of the monolingual model, which is consistent with the complexity analysis in Section 3.

## 5 Conclusion

This paper is devoted to large-scale Chinese UWS for SMT. An efficient unified monolingual and bilingual UWS method is proposed and applied to large-scale bilingual corpora.

Complexity analysis shows that our method is capable of scaling to large-scale corpora. This was verified by experiments on a corpus of 1-million sentence pairs on which traditional MCMC approaches would struggle (Xu et al., 2008).

The proposed method does not require any annotated data, but the SMT system with it can achieve comparable performance compared to state-of-the-art supervised word segmenters trained on precious annotated data. Moreover, the proposed method yields 0.96 BLEU improvement relative to supervised word segmenters on an out-of-domain corpus. Thus, we believe that the proposed method would benefit SMT related to low-resource languages where annotated data are scare, and would also find application in domains that differ too greatly from the domains on which supervised word segmenters were trained.

In future research, we plan to improve the bilingual UWS through applying VB and integrating more accurate alignment models such as HMM models and IBM model 4.

## References


Michael R Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):263–311.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 224–232. Association for Computational Linguistics.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 718–726. Association for Computational Linguistics.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, volume 133.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680. Association for Computational Linguistics.

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.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Brian MacWhinney and Catherine Snow. 1985. The child language data exchange system. *Journal of child language*, 12(2):271–296.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. 2010. Learning a language model from continuous speech. In *InterSpeech*, pages 1053–1056.

ThuyLinh Nguyen, Stephan Vogel, and Noah A Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 815–823. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Michael Paul. 2008. Overview of the IWSLT 2008 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17.

Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Yee Whye Teh and Michael I Jordan. 2010. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics: Principles and Practice*, pages 158–207.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting on Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for SIGHAN Bakeoff 2005. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, volume 171. Jeju Island, Korea.

Anand Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1017–1024. Association for Computational Linguistics.

Hai Zhao and Chunyu Kit. 2008. An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 9–16.