A Simple and Effective Unsupervised Word Segmentation Approach

Songjian Chen
School of Information Science and Technology
Sun Yat-sen University
Guangzhou, 510006, China
csjcg2@gmail.com

Yabo Xu * and **Huiyou Chang**
School of Software
Sun Yat-sen University
Guangzhou, 510006, China
{*xuyabo,isschy*}*@mail.sysu.edu.cn*

Abstract

In this paper, we propose a new unsupervised approach for word segmentation. The core idea of our approach is a novel word induction criterion called WordRank, which estimates the goodness of word hypotheses (character or phoneme sequences). We devise a method to derive exterior word boundary information from the link structures of adjacent word hypotheses and incorporate interior word boundary information to complete the model. In light of WordRank, word segmentation can be modeled as an optimization problem. A Viterbi-styled algorithm is developed for the search of the optimal segmentation. Extensive experiments conducted on phonetic transcripts as well as standard Chinese and Japanese data sets demonstrate the effectiveness of our approach. On the standard Brent version of Bernstein-Ratner corpora, our approach outperforms the state-of-the-art Bayesian models by more than 3%. Plus, our approach is simpler and more efficient than the Bayesian methods. Consequently, our approach is more suitable for real-world applications.

Introduction

Word segmentation, i.e., identifying word boundaries in continuous speech or text, is raised as the fundamental problem in Natural Language Processing (NLP) for its wide application in speech recognition, information extraction, machine translation, etc..

Supervised methods have reported great results for word segmentation in the literature lately (Wang, Zong, and Su 2010), but their applicability is limited in practice due to their dependence on human efforts. In this paper we focus on unsupervised methods which have been increasingly gaining attention in recent NLP research. For word segmentation, unsupervised methods are of great interest for three reasons. Firstly, they can learn to perform accurate word segmentation given input of any human language with little extra manual effort. In addition, they may give computational explanation on how children segment speech and discover words, starting from a state where they don't know any word knowledge. Furthermore, they may support the domain adaptation of supervised methods.

**Corresponding author: Yabo Xu.

Recently the state-of-the-art results are reported by nonparametric Bayesian methods (Goldwater, Griffiths, and Johnson 2006; Johnson 2008; Mochihashi, Yamada, and Ueda 2009). But owing to their inference procedure, they all suffer the problem of high computational cost, which impedes them to be applied in practical applications.

Zhikov (2010) pointed out the problem and presented an efficient algorithm, yet with a little sacrifice in performance. He also pointed out another difficulty of unsupervised word segmentation is to find a reliable word induction criterion.

In this paper, we address both these issues by proposing a sound criterion for word segmentation, which could be computed efficiently.

It is intuitive that word boundary is key to recognize words. Previous works have been proposed to utilize local statistics along with heuristics to infer word boundaries (Sun, Shen, and Tsou 1998; Jin and Tanaka-Ishii 2006). But they focused on deciding the position of word boundaries instead of recognize words as a whole. In contrast, we present a framework for measuring the goodness of word hypotheses in an indirect way in terms of boundary information. It involves two constituents, i.e., exterior boundary values (left-side and right-side) and interior boundary value, which are scores associated to word hypotheses. Firstly we construct link structures of adjacent word hypotheses in order to explore the word boundaries and their relevance. Then we use a link analysis algorithm to calculate the exterior boundary values which measure the goodness of their boundaries. At last we introduce interior boundary value which represents the interior combing degree of word hypotheses, to complete the model. With the criterion, a Viterbi-styled algorithm is developed to search for the optimal segmentation. We conduct extensive experiments on various standard corpus in different languages and our method delivers remarkable performance. On the Brent version of Brent-Ratner corpus (Brent 1999), our method outperforms the state-of-the-art methods using nonparametric Bayesian models. In addition, it is shown that our method is more efficient compared to them.

The rest of of paper is organized as follows. After discussing related work, we describe WordRank as our methodology. Then we describe the experiments and finally conclude the paper.

Related Work

Unsupervised word segmentation are of great interest to NLP researchers. A good number of methods have been proposed in the literature, with fairly good performances reported. To conclude, there are two major categories, i.e., boundary prediction and word recognition.

Boundary prediction methods usually utilize local statistics along with heuristics to decide whether there is a word boundary between two language units (characters, phonemes or syllables) given the local context. The representative examples involve Ando-Lee Criterion (Ando and Lee 2000), Mutual Information (MI) (Sun, Shen, and Tsou 1998) and Branching Entropy (BE) (Jin and Tanaka-Ishii 2006). Recently Fleck (2008) proposed a promising algorithm called WordEnds. It trained a boundary classifier with the utterance boundary cues and then used it to mark word boundaries. Zhikov, Takamura, and Okumura (2010) proposed an efficient algorithm combining the strength of Minimum Description Length (MDL) criterion and local statistics BE. High performance in terms of both accuracy and speed was reported.

In contrast, word recognition methods concentrate on recognizing word units. One class of word recognition methods are based on word induction criteria. They utilize statistic measurements to represent the goodness of word hypotheses and then perform optimal search by virtue of them. Description Length Gain (DLG) (Kit and Wilks 1999) and Accessory Variety (AV) (Feng et al. 2004) fall into this category. Another class refers to the language models. Brent (1999) and Venkataraman (2001) proposed generative models and used incremental search procedure to find the most probable segmentation of the whole corpus. Goldwater, Griffiths, and Johnson (2006) presented an alternative framework with nonparametric Bayesian methods. They developed unigram and bigram models with Dirichlet Process (DP) and Hierarchical Dirichlet Process (HDP) respectively. A Gibbs Sampling algorithm was used for inference. Mochihashi, Yamada, and Ueda (2009) extended their work by introducing a nested model on word spelling and proposing a more efficient sampling procedure. Johnson (2008) presented Adaptor Grammar, a grammar-based framework combining the strength of nonparametric Bayesian methods and various of grammars. The state-of-the-art results are reported by the Bayesian models.

WordRank

In this section we propose a new word induction criterion called WordRank. The intuition behind is that word boundaries between adjacent words indicate the correctness of each other, i.e., if a word hypothesis has a correct (or wrong) word boundary, we may infer that its neighbor would simultaneously have correct (or wrong) word boundary at its corresponding side. It further indicates that the goodness of a word's boundaries depend on their neighbors' boundaries, which is similar to PageRank (Brin and Page 1998) where the importance of a web page depends on all the pages that link to it.

It inspires us to construct link structures based on the adjacent relationship of word hypotheses and use a link analysis algorithm, similar to the HITS algorithm originally proposed for ranking web pages (Kleinberg 1999), to calculate the goodness of word boundaries called Exterior Boundary Value (EBV). And then we estimate the goodness of word hypotheses in terms of EBV.

Constructing the Link Structures

Given an unsegmented corpora, we may retrieve "crude" word hypotheses by considering all the character sequences as word hypotheses. In order to reduce the number of word hypotheses and retrieve valid ones, we use three filtering strategies. Firstly, a practical maximum length of word hypotheses L_{max} is introduced. Then we filter out all the word hypotheses that only occur once in the corpora. Last but not least, we perform an efficient statistical sub-string reduction (Lü, Zhang, and Hu 2004) algorithm (linear time complexity) to remove all the equally frequent character sequences based on the observation that two overlapping character sequences with the same frequence, the shorter one is probably redundant and hence could be discarded. Figure 1(a) depicts a small corpus of three utterances and some retrieved valid word hypotheses along with frequences. [1]

In addition, the retrieving process offers an opportunity to perform filtering operation based on word form constraint. Word hypotheses that violate the constraints are not considered to be valid. Note different constrains are applied according to specific language settings. In this paper we only consider the vowel constraint for English transcripts, i.e., every word must contain a vowel. Whereas for Chinese and Japanese we do not use any word form constraint.

We then construct the link structures based on the adjacent relationship of all the valid word hypotheses. It could be viewed as a graph $G = (V, E_{LN}, E_{RN})$, comprising a set of nodes V and two sets of directed edges E_{LN} and E_{RN}. The nodes represent all the valid word hypotheses. A edge $(p, q) \in E_{LN}$ denotes p occurs on the **left** of q in the corpora, and a edge $(p, q) \in E_{RN}$ denotes p occurs on the **right** of q. Note for symmetry, a edge $(p, q) \in E_{LN}$ indicates a edge $(q, p) \in E_{RN}$, we retain the redundancy in the illustration for clarity and perception.

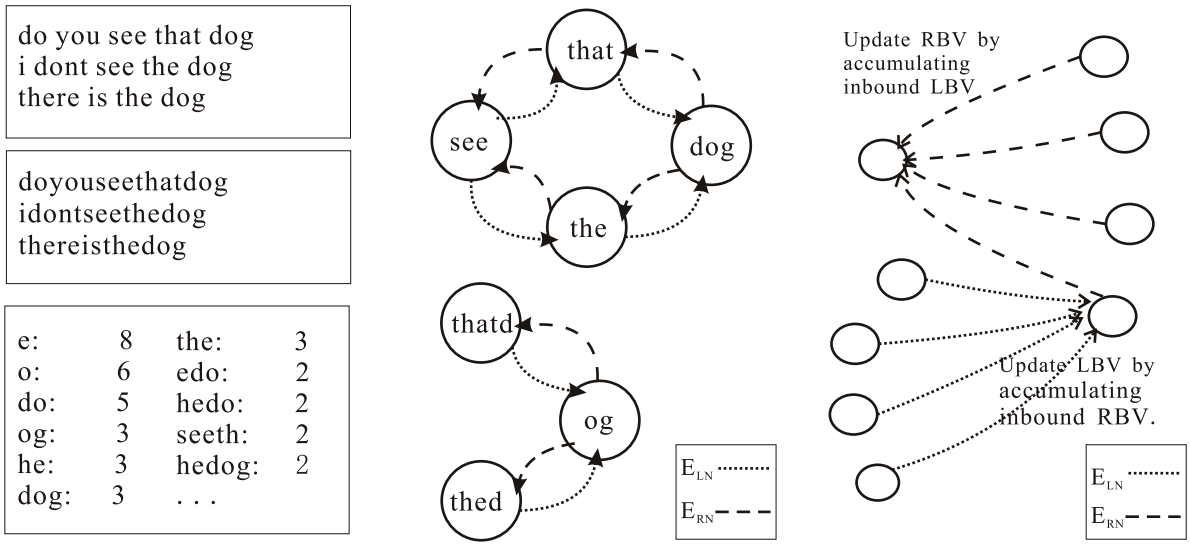
Figure 1(b) shows part of the constructed graph on the corpus of Figure 1(a). Consider "dog" for example, it occurs on the right of "that" in the utterance "doyouseethatdog", as well as "the" in "idontseethedog". Accordingly, there are directed E_{RN} edges from "dog" to "that" and "the" respectively, meanwhile two E_{LN} edges in opposite direction connect to "dog".

Calculating Exterior Boundary Values

Having constructed the link structures of adjacent relationship of word hypotheses, we are allowed to clarify our first observation on EBV of word hypotheses:

Observation 1 *A word hypothesis is likely to have correct left-side word boundary in case it has many left neighbors*

[1] Note we do not perform filtering for the small corpus for the purpose of demonstration.



(a) Segmented, unsegmented corpus and valid word hypotheses

(b) Illustration of the link structure (partial)

(c) Illustration of iterative updating

Figure 1: Illustrations of constructing the link structures of word hypotheses and calculating the exterior boundary values

having right-side word boundaries identified to be correct, and vice versa.

Accordingly, within the link structures described above, boundaries between adjacent word hypotheses exhibit *mutually reinforcing relationship* (Kleinberg 1999). For example, consider word hypotheses occur on the left of "og" in the corpus, "thatd" and "thed" are there, as shown in Figure 1(b). Since "og" has wrong left-side word boundary, both "thatd" and "thed" would hardly have correct right-side word boundary. In other words, the boundary correctness of "og" affects its neighbors, and the "influence" spreads out via the link structure of word hypotheses circularly.

We define two EBVs, i.e., Left-side Boundary Value (LBV) and Right-side Boundary Value (RBV), representing the goodness of left-side and right-side boundary respectively. According to the aforementioned statement, a word hypothesis would receive a high LBV if its left adjacent word hypotheses having high RBVs, and receive high RBV if its right adjacent word hypotheses having high LBVs.

An iterative algorithm is developed for calculating these values. In the initial stage, LBVs and RBVs of all valid word hypotheses are set to 1. At each iteration, the associate values are updated as follows:

$$LBV(w)^{(i+1)} = \sum_{(l,w)\in E_{LN}} RBV(l)^{(i)} \quad (1)$$

$$RBV(w)^{(i+1)} = \sum_{(r,w)\in E_{RN}} LBV(r)^{(i+1)} \quad (2)$$

where E_{LN} and E_{RN} are edge sets of G described above. We update LBVs and RBVs by accumulating the neighboring RBVs and LBVs via inbound edges respectively (See Figure 1(c)). We firstly perform the updating of LBVs of all word hypotheses, then the updating of RBVs.

At each iteration the values are normalized in order to maintain the number scale, by dividing each LBV score by the sum of the squares of all LBV scores, and dividing each RBV score by the sum of the squares of all RBV scores, as follows:

$$LBV(w)' = \frac{LBV(w)}{\sqrt{\sum_t LBV(t)^2}} \quad (3)$$

$$RBV(w)' = \frac{RBV(w)}{\sqrt{\sum_t RBV(t)^2}} \quad (4)$$

We iterate the updating-maintaining procedure for k times to retrieve the convergent scores (The convergence was proved by Kleinberg (1999)). Algorithm 1 presents the specific algorithm.

Consequently we calculate the EBV scores of word hypotheses via combining their LBVs and RBVs as follows [2]:

$$EBV(w) = LBV(w) * RBV(w) \quad (5)$$

Completing with Interior Boundary Value

It is not sound enough to represent the goodness of word hypotheses using only EBV (based on LBV and RBV) since we observe that word combinations (e.g., collocations) would have high LBV and RBV scores for their correct left

[2] We have tried various combining strategies such as addition, minimum and multiplication. The difference is trivial.

Algorithm 1 Iterative Calculating for LBV and RBV

Require: $G(S, E_{LN}, E_{RN})$, n, k
initialize $LBV^{(0)}$, $RBV^{(0)}$
for $i \leftarrow 1$ to k **do**
  for $j \leftarrow 1$ to n **do**
    update $LBV(S[j])^{(i)}$ according to Eq.(1)
  end for
  for $j \leftarrow 1$ to n **do**
    update $RBV(S[j])^{(i)}$ according to Eq.(2)
  end for
  for $j \leftarrow 1$ to n **do**
    normalize $LBV(S[j])^{(i)}$ according to Eq.(3)
  end for
  for $j \leftarrow 1$ to n **do**
    normalize $RBV(S[j])^{(i)}$ according to Eq.(4)
  end for
end for
return ($LBV^{(k)}$,$RBV^{(k)}$)

and right boundary, which leads to a long word bias problem that word combinations are preferred while searching for the optimal segmentation.

The root of this problem is that the exterior boundary values (LBV and RBV) only focus on the exterior of word hypotheses with the interior ignored. We may supplement the deficiency by incorporating a subtle indicator of boundary inside word hypotheses, named Interior Boundary Value (IBV), based on the following observation:

Observation 2 *There are usually strong indicators of word boundary inside word combinations while rarely in true words.*

Consider the following example: "thatdog", is a word hypothesis composed of "that" and "dog". We may easily locate the word boundary between "t" and "d" since "td" hardly occurs within a word in English. In contrast, we may not find this boundary indicator inside a true word such as"that" and "dog".

There are numbers of statistical measurements for the purpose of word boundary indicating (See Section 2 for details). For simplicity of implementation we choose Mutual Information (MI), a well-defined statistical measurement, measuring the combining degree of pairs of adjacent characters (Sun, Shen, and Tsou 1998). It is defined as follows:

$$MI(x:y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

where $p(x, y)$ is the co-occurrence probability of character x and y, and $p(\cdot)$ is the probability that the single character occurs in the corpus. It has been reported that high MI value indicates a good chance of two characters combining together, while low MI value, conversely, indicates a word boundary between the two characters.

Given a word hypothesis with length L, it has $L-1$ combining points, and thus $L-1$ MI values. We use the minimum one to calculate IBV since the lowest MI value is the strongest indicator of word boundary. Another intuition behind is: in terms of MI, we may also consider IBV as the overall interior combining degree of the whole word hypothesis, which depends on the combining degree of its weakest combining point. This is similar to the well-known *Cannikin Law* that how much water a cannikin may hold depends on its shortest plank.

As discussed above, $IBV(w)$ of word hypothesis w is calculated as follows:

$$IBV(w) = min_{i=1}^{L-1}\big(MI(c_i : c_{i+1})\big) \tag{6}$$

where L is the length of w, c_i is the ith character of w.

Finally we calculate WordRank by incorporating EBV and IBV as follows:

$$\begin{aligned} WR(w) &= EBV(w) * f\big(IBV(w)\big) \\ &= LBV(w) * RBV(w) * f\big(IBV(w)\big) \end{aligned} \tag{7}$$

where $f(\cdot)$ is the auxiliary function utilized for searching for the optimal performance. We use two common functions, namely, polynomial and exponential, which are the most representative types of functions, denoted as f_{poly} and f_{exp}:

$$f_{poly}(x) = x^{\alpha} \qquad f_{exp}(x) = \beta^{x}$$

The parameters α and β could be regard as adjustors for balancing the effects of exterior and interior information where higher values indicate higher dominance of interior against exterior. The best state of WordRank is reached when the exterior and interior are balanced. We decide the optimal value of the parameters experimentally.

The factors in Equation (7) represent three parts of information of a word hypothesis respectively, i.e., exterior (left-side and right-side) word boundary information and interior word boundary information. It shows that a word hypothesis with high score of correct exterior boundaries and high score of interior combing degree would be likely to be a correct word, which is fully aligned with the common observation. In this case, the word combinations with high LBV and RBV but low IBV would be less preferred, thus the long word bias problem described above is fixed. In light of this equation, the model of WordRank is completed.

Segmentation Algorithm

For segmentation, we develop a Viterbi-styled algorithm for the search of the optimal segmentation with WordRank scores. Formally, we search for the word segmentation that maximizes the following objective function:

$$S^* = argmax_{w_1 \oplus \ldots \oplus w_n = u} \prod_{i=1}^{n} WR(w_i)$$

where u is an utterance of continuous characters, $w_1 \oplus \ldots \oplus w_n$ is the possible segmentation of u where w_i is some valid word hypothesis. $WR(w)$ is the associate WordRank score of w. Since we regard word segmentation as an optimal problem, the segmentation with highest function value would be considered as the resulting solution.

Experiments

Experiment Setup

In order to demonstrate the effectiveness of our approach, we conduct experiments on the Brent version Bernstein-Ratner corpus (Brent 1999), as well as standard Chinese data sets from SIGHAN Bakeoff-2005 (Emerson 2005) and the Japanese Kyoto corpus (Kurohashi and Nagao 1998). Since most recently proposed methods are evaluated against these corpus, we are allowed to perform comprehensive comparisons.

We firstly extract valid word hypotheses, then construct the link structures of adjacent relationship of word hypotheses and finally calculate the WordRank scores following the method described in Section 3. We set $L_{max} = 11$ for English transcripts and set $L_{max} = 5$ for Chinese and Japanese (L_{max} is determined by the max length of the majority of words observed in the evaluation corpus). In practice we run the updating-maintaining procedure for 30 iterations [3] to compute the exterior boundary values. For computing interior boundary value, we use both polynomial and exponential functions, correspondingly the resulting models are denoted as WR(poly) and WR(exp) respectively. They are incorporated with the Viterbi-styled algorithm proposed in the previous section for searching for the optimal segmentation.

We use word token Precision (P), Recall (R) and F-score (F) as evaluation metrics. All experiments were conducted on the Linux platform, with a 2.4GHz Xeon 3430 CPU and 4GB of memory.

Experiments on English Phonetic Transcripts

We firstly perform experiments on the Bernstein-Ratner corpus for the purpose of directly comparing with the state-of-the-art results. The original corpus of orthographic form has been converted to phonemic representation using a phonemic dictionary with 1-character phone symbols, e.g, the corresponding phonemic representation of "look at this" is "lUk &t DIs".

Recall that our approach involves a single parameter, i.e., α or β, which are parameters of the auxiliary functions for computing IBV. For English transcripts, we experimentally use $\alpha = 4.4$ and $\beta = 4.6$. Since the parameter value solely counts on characteristics of the target languages, e.g., morpheme complexity, average word length, etc., we only need a small annotated development data to decide the optimal parameter value while adapting our method to different languages.

In Table 1, we compare the results of our method to that of previously proposed models (We collect the results from the literatures). Notably our method (either with polynomial or exponential auxiliary function) achieves a 78% F-score, outperforming all the previous results. First of all, it exemplifies WordRank is a sound word induction criterion for word segmentation. Due to the deliberate modeling of both exterior and interior information, it avoids the problem that other word induction criterion based methods may suffer, e.g., prefer collocations or single character as a word. Our method also outperforms the state-of-the-art Bayesian models, confirming the effectiveness of the approach. Besides, it is shown that our method is superior in terms of efficiency. Table 2 shows the comparison of computational time with Bayesian models. Our learner only needs around one fifth of the running time of NPY (Mochihashi, Yamada, and Ueda 2009) which used an improved inference procedure. We find the iterative procedure for computing WordRank scores are the most time consuming step (around 83% of the total time). As a matter of fact, it could be remarkably improved through parallel computing, as the way Google computes their page ranks. We leave this for future work.

Models	P (%)	R (%)	F (%)
WordEnds	-	-	70.7
Ent-MDL	-	-	75.4
HDP(2)	75.2	69.6	72.3
NPY(2)	74.8	76.7	75.7
NPY(3)	74.8	75.2	75.0
WR (poly)	**78.8**	78.5	**78.6**
WR (exp)	77.6	**78.6**	78.1

Table 1: Comparison of our method with previously proposed models, i.e., WordEns (Fleck 2008) and Ent-MDL (Zhikov, Takamura, and Okumura 2010), as well as nonparametric Bayesian models, i.e., HDP (Goldwater, Griffiths, and Mark 2009) and NPY (Mochihashi, Yamada, and Ueda 2009) on English phonetic transcripts. The n (in parentheses) means n-gram model

Models	Time
HDP	10h55min
NPY	17min
WR	<**3min**

Table 2: Comparison of computation time of our method with Bayesian models

Experiments on Chinese and Japanese Data Sets

Extended experiments are conducted on Chinese and Japanese data sets to demonstrate the applicability of our method to other languages. We decide the auxiliary parameter experimentally in the same way as for English transcripts.[4]

Approaches	MSR	CityU	Kyoto
NPY(2)	80.2	82.4	62.1
NPY(3)	80.7	81.7	66.6
WR (poly)	78.6	78.7	65.0
WR (exp)	79.3	78.8	66.3

Table 3: Comparison of our method with nonparametric Bayesian models on Chinese and Japanese data sets

Table 3 sums up the comparison of our method with the

[3] Actually the scores approximately converge around 10 iterations.

[4] Within the experiments we use $\alpha = 5.5$ and $\beta = 3.4$ for Chinese, while $\alpha = 5.0$ and $\beta = 3.2$ for Japanese.

state-of-the-art models. The Chinese results are similar to English results, confirming that WordRank is a sound word induction criterion for different languages with distinct morphological structures. Whereas Japanese F-score appears a little low. We may not able to propose an error analysis owing to our poor knowledge on Japanese, yet the state-of-the-art nonparametric Bayesian models present similar accuracy as shown in Table 3. It should be noticed that Mochihashi, Yamada, and Ueda (2009) used separate testing corpus of Chinese data sets and random subset of the Kyoto corpus for evaluation, thus the comparison is not direct.

Conclusion

In this paper, we proposed a simple and effective unsupervised word segmentation approach. We introduced a novel word induction criterion called WordRank, for measuring the goodness of word hypotheses. The criterion incorporates both the exterior and interior boundary information to model words. We devise a method to derive exterior boundary value from the link structures of adjacent word hypotheses and incorporate inner boundary value to complete the model. A Viterbi-styled algorithm is developed for the search of the optimal segmentation. Extensive experiments confirm the soundness of our proposed word induction criterion for word segmentation. It is also exemplified by the experiments that our word segmentation approach based on the proposed criterion is simple, efficient and effective. Thus it is a suitable method for real-word applications.

Acknowledgements

The research of Yabo Xu is supported by Sun Yat-sen University Grants 62000-3161032 and 62000-3165002. We would also like to thank Sharon Goldwater, Margaret M. Fleck, Daichi Mochihashi and Valentin Zhikov for sharing the data sets, and the anonymous reviewers for suggestions.

References

Ando, R. K., and Lee, L. 2000. Mostly-unsupervised statistical segmentation of Japanese: applications to kanji. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 241–248.

Brent, M. R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* 34:71–105.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web*, 107–117.

Emerson, T. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Feng, H.; Chen, K.; Kit, C.; and Deng, X. 2004. Unsupervised segmentation of Chinese corpus using accessor variety. In Su, K.-Y.; Tsujii, J.; Lee, J.-H.; and Kwong, O., eds., *Natural Language Processing IJCNLP 2004*, volume 3248 of *Lecture Notes in Computer Science*. 694–703.

Fleck, M. M. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of the Association for Computational Linguistics (ACL)*, 130–138.

Goldwater, S.; Griffiths, T. L.; and Johnson, M. 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*, 673–680.

Goldwater, S.; Griffiths, T. L.; and Mark, J. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1):21 – 54.

Jin, Z., and Tanaka-Ishii, K. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 428–435.

Johnson, M. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of ACL-08: HLT*, 398–406.

Kit, C., and Wilks, Y. 1999. Unsupervised learning of word boundary with description length gain. In *Proceedings of CoNLL*, 1–6.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of ACM* 46:604–632.

Kurohashi, S., and Nagao, M. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceeding of The First International Conference on Language Resources and Evaluation*, 719–724.

Lü, X.; Zhang, L.; and Hu, J. 2004. Statistical substring reduction in linear time. In Su, K.-Y.; Tsujii, J.; Lee, J.-H.; and Kwong, O., eds., *Natural Language Processing IJCNLP 2004*, volume 3248. 320–327.

Mochihashi, D.; Yamada, T.; and Ueda, N. 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*, 100–108.

Sun, M.; Shen, D.; and Tsou, B. K. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, 1265–1271.

Venkataraman, A. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics* 27:352–372.

Wang, K.; Zong, C.; and Su, K.-Y. 2010. A character-based joint model for Chinese word segmentation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 1173–1181.

Zhikov, V.; Takamura, H.; and Okumura, M. 2010. An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 832–842.