A Joint Model for Unsupervised Chinese Word Segmentation

Miaohong Chen Baobao Chang Wenzhe Pei
Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
Beijing, P.R.China, 100871
miaohong-chen@foxmail.com,{chbb,peiwenzhe}@pku.edu.cn

Abstract

In this paper, we propose a joint model for unsupervised Chinese word segmentation (CWS). Inspired by the "products of experts" idea, our joint model firstly combines two generative models, which are word-based hierarchical Dirichlet process model and character-based hidden Markov model, by simply multiplying their probabilities together. Gibbs sampling is used for model inference. In order to further combine the strength of goodness-based model, we then integrated nVBE into our joint model by using it to initializing the Gibbs sampler. We conduct our experiments on PKU and MSRA datasets provided by the second SIGHAN bakeoff. Test results on these two datasets show that the joint model achieves much better results than all of its component models. Statistical significance tests also show that it is significantly better than state-of-the-art systems, achieving the highest F-scores. Finally, analysis indicates that compared with nVBE and HDP, the joint model has a stronger ability to solve both combinational and overlapping ambiguities in Chinese word segmentation.

1 Introduction

Unlike English and many other western languages, there are no explicit word boundaries in Chinese sentences. Therefore, word segmentation is a crucial first step for many Chinese language processing tasks such as syntactic parsing, information retrieval and machine translation. A great deal of supervised methods have been proposed for Chinese word segmentation. While successful, they require manually labeled resources and often suffer from issues like poor domain adaptability. Thus, unsupervised word segmentation methods are still attractive to researchers due to its independence on domain and manually labeled corpora.

Previous unsupervised approaches to word segmentation can be roughly classified into two types. The first type uses carefully designed goodness measure to identify word candidates. Popular goodness measures include description length gain (DLG) (Kit and Wilks, 1999), accessor variety (AV) (Feng et al., 2004), boundary entropy (BE) (Jin and Tanaka-Ishii, 2006) and normalized variation of branching entropy (nVBE) (Magistry and Sagot, 2012) etc. Goodness measure based model is not segmentation model in a very strict meaning and is actually strong in generating word list without supervision. It inherently lacks capability to deal with ambiguous string, which is one of main sources of segmentation errors and has been extensively explored in supervised Chinese word segmentation.

The second type focuses on designing sophisticated statistical model, usually nonparametric Bayesian models, to find the segmentation with highest posterior probability, given the observed character sequences. Typical statistical models includes Hierarchical Dirichlet process (HDP) model (Goldwater et al., 2009), Nested Pitman-Yor process (NPY) model (Mochihashi et al., 2009) etc, which are actually nonparametric language models and therefor can be categorized as word-based model. Word-based model makes decision on wordhood of a candidate character sequence mainly based on information outside the sequence, namely, the wordhood of character sequences being adjacent to the concerned sequence.

Inspired by the success of character-based model in supervised word segmentation, we propose a Bayesian HMM model for unsupervised Chinese word segmentation. With the Bayesian HMM model, we formulate the unsupervised segmentation tasks as procedure of tagging positional

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 854–863,
October 25-29, 2014, Doha, Qatar.

tags to characters. Different from word-based model, character-based model like HMM-based model as we propose make decisions on wordhood of a candidate character sequence based on information inside the sequence, namely, ability of characters to form words. Although the Bayesian HMM model alone does not produce competitive results, it contributes substantially to the joint model as proposed in this paper.

Our joint model takes advantage from three different models: namely, a character-based model (HMM-based), a word-based model (HDP-based) and a goodness measure based model (nVBE model). The combination of HDP-based model and HMM-based model enables to utilize information of both word-level and character-level. We also show that using nVBE model as initialization model could further improve the performance to outperform the state-of-the-art systems and leads to improvement in both wordhood judgment and disambiguation ability.

Word segmentation systems are usually evaluated with metrics like precision, recall and F-Score, regardless of supervised or unsupervised. Following normal practice, we evaluate our model and compare it with state-of-the-art systems using F-Score. However, we argue that the ability to solve segmentation ambiguities is also important when evaluating different types of unsupervised word segmentation systems.

This paper is organized as follows. In Section 2, we will introduce several related systems for unsupervised word segmentation. Then our joint model is presented in Section 3. Section 4 shows our experiment results on the benchmark datasets and Section 5 concludes the paper.

2 Related Work

Unsupervised Chinese word segmentation has been explored in a number of previous works and by various methods. Most of these methods can be divided into two categories: goodness measure based methods and nonparametric Bayesian methods.

There have been a plenty of work that is based on a specific goodness measure. Zhao and Kit (2008) compared several popular unsupervised models within a unified framework. They tried various types of goodness measures, such as Description Length Gain (DLG) proposed by Kit and Wilks (1999), Accessor Variety (AV) proposed by Feng et al. (2004) and Boundary Entropy (Jin and Tanaka-Ishii, 2006). A notable goodness-based method is ESA: "Evaluation, Selection, Adjustment", which is proposed by Wang et al. (2011) for unsupervised Mandarin Chinese word segmentation. ESA is an iterative model based on a new goodness algorithm that adopts a local maximum strategy and avoids threshold setting. One disadvantage of ESA is that it needs to iterate the process several times on the corpus to get good performance. Another disadvantage is the requirement for a manually segmented training corpus to find best value for parameters (they called it *proper exponent*). Another notable work is nVBE: Magistry and Sagot (2012) proposed a model based on the Variation of Branching Entropy. By adding normalization and viterbi decoding, they improve performance over Jin and Tanaka-Ishii (2006) and remove most of the parameters and thresholds from the model.

Nonparametric Bayesian models also achieved state-of-the-art performance in unsupervised word segmentation. Goldwater et al. (2009) introduced a unigram and a bigram model for unsupervised word segmentation, which are based on Dirichlet process and hierarchical Dirichlet process (Teh et al., 2006) respectively. The main drawback is that it needs almost 20,000 iterations before the Gibbs sampler converges. Mochihashi et al. (2009) extended this method by introducing a nested character model and an efficient blocked Gibbs sampler. Their method is based on what they called nested Pitman-Yor language model.

One disadvantage of goodness measure based methods is that they do not have any disambiguation ability in theory in spite of their competitive performances. This is because once the goodness measure is given, the decoding algorithm will segment any ambiguous strings into the same word sequences, no matter what their context is. In contrast, nonparametric Bayesian language models aim to segment character string into a "reasonable" sentence according to the posterior probability. Thus, theoretically, this method should have better ability to solve ambiguities over goodness measure based methods.

3 Joint Model

In this section, we will discuss our joint model in detail.

3.1 Combining HDP and HMM

In supervised Chinese word segmentation literature, word-based approaches and character-based approaches often have complementary advantages (Wang et al., 2010).Since the two types of model try to solve the problem from different perspectives and by utilizing different levels of information (word level and character level). In unsupervised Chinese word segmentation literature, the HDP-base model can be viewed as a typical word-based method. And we can also build a character-based unsupervised model by using a hidden Markov model. We believe that the HDP-based model and the HMM-based model are also complementary with each other, and a combination of them will take advantage of both and thus capture different levels of information.

Now the problem we are facing is how to combine these two models. To keep the joint model simple and involve as little extra parameters as possible, we combine the two baseline models by just multiplying their probabilities together and then renormalizing it. Let $C = c_1 c_2 \cdots c_{|C|}$ be a string of characters and $W = w_1 w_2 \cdots w_{|W|}$ is the corresponding segmented words sequence. Then the conditional probability of the segmentation W given the character string C in our joint model is defined as:

$$P_J(W|C) = \frac{1}{Z(C)} P_D(W|C) P_M(W|C) \quad (1)$$

where $P_D(W|C)$ is the probability from the HDP model as given in Equation 6 and $P_M(W|C)$ is the probability given by the Bayesian HMM model as given in Equation 2. $Z(C)$ is a normalization term to make sure that $P_J(W|C)$ is a probability distribution. The combining method is inspired by Hinton (1999), which proved that it is possible to combine many individual expert models by multiplying the probabilities and then renormalizing it. They called it "product of experts". We can see that combining models in this way does not involve any extra parameters and Gibbs sampling can be easily used for model inference.

3.2 Bayesian HMM

The dominant method for supervised Chinese word segmentation is character-based model which was first proposed by Xue (2003). This method treats word segmentation as a tagging problem, each tag indicates the position of a character within a word. The most commonly used tag set is {**S**ingle, **B**egin, **M**iddle, **E**nd}. Specifically, **S** means the character forms a single word, **B/E** means the character is the begining/ending character of the word, and **M** means the character is in the middle of the word. Existing models are trained on manually annotated data in a supervised way based on discriminative models such as Conditional Random Fields (Peng et al., 2004; Tseng et al., 2005). Supervised character-based methods make full use of character level information and thus have been very successful in the last decade. However, no unsupervised model has utilized character level information in the way as supervised method does.

We can also build a character-based model for Chinese word segmentation using hidden Markov model(HMM) as formulated in the following equation:

$$P_M(W|C) = \prod_{i=1}^{|C|} P_t(t_i|t_{i-1}) P_e(c_i|t_i) \quad (2)$$

where C and W have the same meaning as before. $P_t(t_i|t_{i-1})$ is the transition probability of tag t_i given its former tag t_{i-1} and $P_e(c_i|t_i)$ is the emission probability of character c_i given its tag t_i. This model can be easily trained with Maximum Likelihood Estimation (MLE) on annotated data or with Expectation Maximization (EM) on raw texts. But using any of this methods will make it difficult to combine it with the HDP-based model. Instead, we propose a Bayesian HMM for unsupervised word segmentation. The Bayesian HMM model is defined as follows:

$$\begin{aligned} t_i|t_{i-1} = t, p^t &\sim Mult(p^t) \\ c_i|t_i = t, e^t &\sim Mult(e^t) \\ p^t|\theta &\sim Dirichlet(\theta) \\ e^t|\sigma &\sim Dirichlet(\sigma) \end{aligned}$$

where p^t and e^t are transition and emission distributions, θ and σ are the symmetric parameters of Dirichlet distributions. Now suppose we have observed tagged text h, then the conditional probability $P_M(w_i|w_{i-1} = l, h)$ can be obtained:

$$\begin{aligned} &P_M(w_i|w_{i-1} = l, h) \\ &\quad = \prod_{j=1}^{|w_i|} P_t(t_j|t_{j-1}, h) P_e(c_j|t_j, h) \quad (3) \end{aligned}$$

where $< w_{i-1}, w_i >$ is a word bigram, l is the index of word w_{i-1}, c_j is the jth character in word

w_i and t_j is the corresponding tag.$P_t(t_j|t_{j-1}, h)$ and $P_e(c_j|t_j, h)$ are the posterior probabilities, they are given as:

$$P_t(t_j|t_{j-1}, h) = \frac{n_{<t_{j-1},t_j>} + \theta}{n_{<t_{j-1},*>} + T\theta} \quad (4)$$

$$P_e(c_j|t_j, h) = \frac{n_{<t_j,c_j>} + \sigma}{n_{<t_j,*>} + V\sigma} \quad (5)$$

where $n_{<t_{j-1},t_j>}$ is the tag bigram count of $< t_{j-1}, t_j >$ in h, $n_{<t_j,c_j>}$ denotes the number of occurrences of tag t_j and character c_j, and $*$ means a sum operation. T and V are the size of character tag set (we follow the commonly used {**SBME**} tag set and thus $T = 4$ in this case) and character vocabulary.

3.3 HDP Model

Goldwater et al. (2009) proposed a nonparametric Bayesian model for unsupervised word segmentation which is based on HDP (Teh et al., 2006). In this model, the conditional probability of the segmentation W given the character string C is defined as:

$$P_D(W|C) = \prod_{i=0}^{|W|} P_D(w_i|w_{i-1}) \quad (6)$$

where w_i is the ith word in W. This is actually a nonparametric bigram language model. This bigram model assumes that each different word has a different distribution over words following it, but all these different distributions are linked through a HDP model:

$$\begin{aligned} w_i|w_{i-1} = l &\sim G_l \\ G_l &\sim DP(\alpha_1, G_0) \\ G_0 &\sim DP(\alpha, H) \end{aligned}$$

where DP denotes a Dirichlet process.

Suppose we have observed segmentation result h, then we can get the posterior probability $P_D(w_i|w_{i-1} = l, h)$ by integrating out G_l:

$$\begin{aligned} &P_D(w_i|w_{i-1} = l, h) \\ &\quad = \frac{n_{<w_{i-1},w_i>} + \alpha_1 P_D(w_i|h)}{n_{<w_{i-1},*>} + \alpha_1} \end{aligned} \quad (7)$$

where $n_{<w_{i-1},w_i>}$ denotes the total number of occurrences of the bigram $< w_{i-1}, w_i >$ in the observation h. And $P_D(w_i|h)$ can be got by integrating out G_0:

$$P_D(w_i|h) = \frac{t_{w_i} + \alpha H(w_i)}{t + \alpha} \quad (8)$$

where t_{w_i} denotes the number of tables associated with w_i in the Chinese Restaurant Franchise metaphor (Teh et al., 2006), t is the total number of tables and $H(w_i)$ is the base measure of G_0. In fact, $H(w_i)$ is the prior distribution over words, so prior knowledge can be injected in this distribution to enhance the performance.

In Goldwater et al. (2009)'s work, the base measure $H(w_i)$ are defined as a character unigram model:

$$H(w_i) = (1 - p_s)^{|w_i|-1} p_s \prod_j P(c_{ij})$$

where, p_s is the probability of generating a word boundary. $P(c_{ij})$ is the probability of the jth character c_{ij} in word w_i, this probability can be estimated from the training data using maximum likelihood estimation.

3.4 Initializing with nVBE

Among various goodness measure based models, we choose nVBE (Magistry and Sagot, 2012) to initialize our Gibbs sampler with its segmentation results. nVBE achieved a relatively high perfomance over other goodness measure based methods. And it's very simple as well as efficient.

Theoretically, the Gibbs sampler may be initialized at random or using any other methods. Initialization does not make a difference since the Gibbs sampler will eventually converge to the posterior distribution if it iterates as much as possible. This is an essential attribute of Gibbs sampling. However, we believe that initializing the Gibbs sampler with the result of nVBE will benefit us in two ways. On one hand, in consideration of its combination of nonparametric Bayesian method and goodness-based method, it will improve the overall performance as well as solve more segmentation ambiguities with the help of HDP-based model. On the other hand, it makes the convergence of Gibbs sampling faster. In practice, random initialization often leads to extremely slow convergence.

3.5 Inference with Gibbs Sampling

In our proposed joint model, Gibbs sampling (Casella and George, 1992) can be easily used to identify the highest probability segmentation from among all possibilities. Following Goldwater et al. (2009), we can repeatedly sample from potential word boundaries. Each boundary

variable can only take on two possible values, corresponding to a word boundary or not word boundary.

For instance, suppose we have obtained a segmentation result $\beta|c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}|\gamma$, where β and γ are the words sequences to the left and right and $c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}$ are characters between them. Now we are sampling at location i to decide whether there is a word boundary between c_i and c_{i+1}. Denote h_1 as the hypothesis that it forms a word boundary (the corresponding result is $\beta w_1 w_2 \gamma$ where $w_1 = c_{i-2}c_{i-1}c_i$ and $w_2 = c_{i+1}c_{i+2}$), and h_2 as the opposite hypothesis (then the corresponding result is $\beta w \gamma$ where $w = c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}$). The posterior probability for these two hypotheses would be:

$$P(h_1|h^-) \propto P_D(h_1|h^-)P_M(h_1|h^-) \quad (9)$$
$$P(h_2|h^-) \propto P_D(h_2|h^-)P_M(h_2|h^-) \quad (10)$$

where $P_D(h|h^-)$ and $P_M(h|h^-)$ are the posterior probabilities in HDP-based model and in HMM-based model, and h^- denotes the current segmentation results for all observed data except $c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}$. Note that the normalization term $Z(C)$ can be ignored during inference. The posterior probabilities for these two hypotheses in the HDP-based model is given as:

$$\begin{aligned} P_D(h_1|h^-) &= P_D(w_1|w_l, h^-) \\ &\times P_D(w_2|w_1, h^-)P_D(w_r|w_2, h^-) \quad (11) \\ P_D(h_2|h^-) &= P_D(w|w_l, h^-) \\ &\times P_D(w_r|w, h^-) \quad (12) \end{aligned}$$

where $w_l(w_r)$ is the first word to the left (right) of w. And the posterior probabilities for the Bayesian HMM model is given as:

$$\begin{aligned} &P_M(h_1|h^-) \\ &\propto \prod_{j=i-2}^{i+2} P_t(t_j|t_{j-1}, h^-)P_e(c_j|t_j, h^-) \quad (13) \\ &P_M(h_2|h^-) \\ &\propto \prod_{j=i-2}^{i+2} P_t(t_j|t_{j-1}, h^-)P_e(c_j|t_j, h^-) \quad (14) \end{aligned}$$

where $P_t(t_j|t_{j-1}, h^-)$ and $P_e(c_j|t_j, h^-)$ are given in Equation 4 and 5. The difference is that under hypothesis h_1, $c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}$ are tagged as "BMEBE" and under hypothesis h_2 as "BMMME".

Once the Gibbs sampler is converged, a natural way to is to treat the result of last iteration as the final segmentation result, since each set of assignments to the boundary variables uniquely determines a segmentation.

4 Experiments

In this section, we test our joint model on PKU and MSRA datesets provided by the Second Segmentation Bake-off (SIGHAN 2005) (Emerson, 2005). Most previous works reported their results on these two datasets, this will make it convenient to directly compare our joint model with theirs.

4.1 Setting

The second SIGHAN Bakeoff provides several large-scale labeled data for evaluating the performance of Chinese word segmentation systems. Two of the four datasets are used in our experiments. Both of the dataset contains only simplified Chinese. Table 1 shows the statistics of the two selected corpus. For development set, we randomly select a small subset (about 10%) of the training data. Specifically, 2000 sentences are selected for PKU corpus and 8000 sentences for MSRA corpus. The rest training data plus the test set is then combined for segmentation but only test data is used for evaluation. The development set is used to tune parameters of the HDP-based model and HMM-based model separately. Since our joint model does not involve any additional parameters, we reuse the parameters of the HDP-based model and HMM-based model in the joint model. Specifically, we set $\alpha_1 = 1000.0, \alpha = 10.0, p_s = 0.5$ for the HDP-based model and set $\theta = 1.0, \sigma = 0.01$ for the HMM-based model.

For evaluation, we use standard F-Score on words for all following experiments. F-Score is the harmonic mean of the word precision and recall. Precision is given as:

$$P = \frac{\#correct\ words\ in\ result}{\#total\ words\ in\ result}$$

and recall is given as:

$$R = \frac{\#correct\ words\ in\ result}{\#total\ words\ in\ gold\ corpus}$$

then F-Score is calculated as:

$$F = \frac{2 \times R \times F}{R + F}$$

Corpus	TrainingSize (words)	TestSize (words)
PKU	1.1M	104K
MSRA	2.37M	107K

Table 1: Statistics of training and testing data

Huang and Zhao (2007) provided an empirical method to estimate the consistency between the four different segmentation standards involved in the Bakeoff-3. A lowest consistency rate 84.8% is found among the four standards. Zhao and Kit (2008) considered this figure as the upper bound for any unsupervised Chinese word segmentation systems. We also use it as the **topline** in our comparison.

4.2 Prior Knowledge Used

When it comes to the evaluation and comparison for unsupervised word segmentation systems, an important issue is what kind of pre-processing steps and prior knowledge are needed. To be fully unsupervised, any prior knowledge such as punctuation information, encoding scheme and word length could not be used in principle. Nevertheless, information like punctuation can be easily injected to most existing systems and significantly enhance the performance. The problem we are faced with is that we don't know for sure what kind of prior information are used in other systems. One may use a small punctuation set to segment a long sentence into shorter ones, while another may write simple regular expressions to identify dates and numbers. Lot of work we compare to don't even mention this subject.

Fortunately, we notice that Wang et al. (2011) provided four kinds of preprocessings (they call *settings*). In their settings 1 and 2, punctuation and other encoding information are not used. In setting 3, punctuation is used to segment character sequences into sentences, and both punctuation and other encoding information are used in setting 4. Then the results reported in Magistry and Sagot (2012) relied on setting 3 and setting 4. In order to make the comparison as fair as possible, we use setting 3 in our experiment, i.e., only a punctuation set for simplified Chinese is used in all our experiments. We will compare our experiment results to previous work on the same setting if they are provided .

4.3 Experiment Results

Table 2 summarizes the F-Scores obtained by different models on PKU and MSRA corpus, as well as several state-of-the-art systems. Detailed information about the presented models are listed as follows:

- **nVBE:** the model based on Variation of Branching Entropy in Magistry and Sagot (2012). We re-implement their model on setting 3[1].
- **HDP:** the HDP-based model proposed by Goldwater et al. (2009), initialized randomly.
- **HDP+HMM:** the model combining HDP-based model and HMM-based model as proposed in Section 3, initialized randomly.
- **HDP+nVBE:** the HDP-based model, initialized with the results of nVBE model.
- **Joint:** the "HDP+HMM" model initialized with nVBE model.
- **ESA:** the model proposed in Wang et al. (2011), as mentioned above, the conducted experiments on four different settings, we report their results on setting 3.
- **NPY(2):** the 2-gram language model presented by Mochihashi et al. (2009).
- **NPY(3):** the 3-gram language model presented by Mochihashi et al. (2009).

For all of our Gibbs samplers, we run 5 times to get the averaged F-Scores. We also give the variance of the F-Scores in Table 2. For each run, we find that random initialization takes around 1,000 iterations to converge, while initialing with nVBE only takes as few as 10 iterations. This makes

[1]The results we got with our implementation is slightly lower than what was reported in Magistry and Sagot (2012). According to Pei et al. (2013), they had contacted the authors and confirmed that the higher results was due to a bug in code. So we report the results with our bug free implementation as Pei et al. (2013) did. Our reported results are identical to those of Pei et al. (2013)

System	PKU			MSRA		
	R	**P**	**F**	**R**	**P**	**F**
nVBE	78.3	77.5	77.9	79.1	77.3	78.2
HDP	69.0	68.4	68.7(0.012)	70.4	69.4	69.9(0.020)
HDP+HMM	77.5	73.2	75.3(0.005)	79.9	73.0	76.3(0.013)
HDP+nVBE	80.7	77.9	79.3(0.012)	81.8	77.3	79.5(0.005)
Joint	**83.1**	**79.2**	**81.1**(0.002)	**84.2**	**79.3**	**81.7**(0.005)
ESA	N/A	N/A	77.4	N/A	N/A	78.4
NPY(2)	N/A	N/A	N/A	N/A	N/A	80.2
NPY(3)	N/A	N/A	N/A	N/A	N/A	80.7
Topline	N/A	N/A	84.8	N/A	N/A	84.8

Table 2: Experiment results and comparison to state-of-the-art systems. The figures in parentheses denote the variance the of F-Scores.

our joint model very efficient and possible to work in practical applications as well. At last, a single sample (the last one) is used for evaluation.

From Table 2, we can see that the joint model (Joint) outperforms all the presented systems in F-Score on all testing corpora. Specifically, comparing "HDP+HMM" with "HDP", the former model increases the overall F-Score from 68.7% to 75.3% (+6.6%) in PKU corpora and from 69.9% to 76.3% (+6.4%) in MSRA corpora, which proves that the character information in the HMM-based model can actually enhance the performance of the HDP-based model. Comparing "HDP+nVBE" with "HDP", the former model also increases the overall F-Score by 10.6%/9.6% in PKU/MSRA corpora, which demonstrates that initializing the HDP-based model with nVBE will improve the performance by a large margin. Finally, the joint model "Joint" take advantage from both from the character-based HMM model and the nVBE model, it achieves a F-Score of 81.1% on PKU and 81.7% on MSRA. This result outperforms all its component baselines such as "HDP", "HDP+HMM" and "HDP+nVBE".

Our joint model also shows competitive advantages over several state-of-the-art systems. Compared with nVBE,the F-Score increases by 3.2% on PKU corpora and by 3.5% on MSRA corpora. Compared with ESA, the F-Score increases by 3.7%/3.3% in PKU/MSRA corpora. Lastly, compared to the nonparametric Bayesian models (NPY(n)), our joint model still increases the F-Score by 1.5% (NPY(2)) and 1.0% (NPY(3)) on MSRA corpora. Moreover, compared with the empirical topline figure 84.8%, our joint model achieves a pretty close F-Score. The differences are 3.7% on PKU corpora and 3.1% on MSRA corpora.

An phenomenon we should pay attention to is the poor performance of the HMM-based model. With our implementation of the Bayesian HMM, we achieves a 34.3% F-Score on PKU corpora and a 34.9% F-Score on MSRA corpora, just slightly better than random segmentation. The result show that the hidden Markov Model alone is not suitable for character-based Chinese word segmentation problem. However, it still substantially contributes to the joint model.

We find that the variance of the results are rather small, this shows the stability of our Gibbs samplers. From the segmentation results generated by the joint model, we also found that quite a large amount of errors it made are related to dates, numbers (both Chinese and English) and English words. This problem can be easily addressed during preprocessing by considering encoding information as previous work, and we believe this will bring us much better performance.

4.4 Disambiguation Ability

Previous unsupervised work usually evaluated their models using F-score, regardless of goodness measure based model or nonparametric Bayesian model. However, segmentation ambiguity is a very important factor influencing accuracy of Chinese word segmentation systems (Huang and Zhao, 2007). We believe that the disambiguation ability of the models should also be considered when evaluating different types of unsupervised segmentation systems, since different type of models shows different disambiguation ability. We will compare the disambiguation ability of dif-

ferent systems in this section.

In general, there are mainly two kinds of ambiguity in Chinese word segmentation problem:

- **Combinational Ambiguity**: Given character strings "A" and "B", if "A", "B", "AB" are all in the vocabulary, and "AB" or "A-B" (here "-" denotes a space) occurred in the real text,then "AB" can be called a combinational ambiguous string.
- **Overlapping Ambiguity**: Given character strings "A", "J" and "B", if "A", "B", "AJ" and "JB" are all in the vocabulary, and "A-JB" or "AJ-B" occurred in the real text, then "AJB" can be called an overlapping ambiguous string.

We count the total number of mistakes different systems made at ambiguous strings (the vocabulary is obtained from the gold standard answer of testing set). As we have mentioned in Section 2, goodness measure based methods such as nVBE do not have any disambiguation ability in theory. Our observation is identical to this argument. We find that nVBE always segments ambiguous strings into the same result. Take a combinational string "只有" as an example, "只 (just)", "有 (have)" and "只有 (only)" are all in the vocabulary. In the PKU test set, this string occurs 14 times as "只-有 (just have)" and 18 times as "只有 (only)", 32 times in total. nVBE segments all the 32 strings into "只有 (only)" (i.e. 18 of them are correct), while the joint model segments it 22 times as "只有 (only)" and 10 times as "只-有 (just have)" according to its context, and 24 of them are correct.

Table 3 and 4 show the statistics of combinational ambiguity and overlapping ambiguity respectively. The numbers in parentheses denote the total number of ambiguous strings. From these tables, we can see that HDP+nVBE makes less mistakes than nVBE in most circumstances, except that it solves less combinational ambiguities on MSRA corpora. But our proposed joint model solves the most combinational and overlapping ambiguities, on both PKU and MSRA corpora. Specifically, compared to nVBE, the joint model correctly solves 171/871 more combinational ambiguities on PKU/MSRA corpora, which is a $0.6\%/13.8\%$ relative error reduction. It also solves 28/45 more overlapping ambiguities on PKU/MSRA corpora, which is a $11.5\%/23.4\%$ relative error reduction. This indicates that the joint model has a stronger ability of disambiguation over the compared systems.

System	PKU(35371)	MSRA(38506)
nVBE	8087	7236
HDP+nVBE	7970	7500
Joint	**7916**	**6305**

Table 3: Statistics of combinational ambiguity. This table shows the total number of mistakes made by different systems at combinational ambiguous strings. The numbers in parentheses denote the total number of combinational ambiguous strings.

System	PKU(603)	MSRA(467)
nVBE	244	192
HDP+nVBE	239	164
Joint	**216**	**157**

Table 4: Statistics of overlapping ambiguity. This table shows the total number of mistakes made by different systems at overlapping ambiguous strings. The numbers in parentheses denote the total number of overlapping ambiguous strings.

4.5 Statistical Significance Test

The main results presented in Table 2 has shown that our proposed joint model outperforms the two baselines as well as state-of-the-art systems. But it is also important to know if the improvement is statistically significant over these systems. So we conduct statistical significance tests of F-scores among these various models. Following Wang et al. (2010), we use the bootstrapping method (Zhang et al., 2004).

Here is how it works: suppose we have a testing set T_0 to test several word segmentation systems, there are N testing examples (sentences or line of characters) in T_0. We create a new testing set T_1 with N examples by sampling with replacement from T_0, then repeat these process $M-1$ times. And we will have a total $M+1$ testing sets. In our test procedures, M is set to 2000.

Since we just implement our joint model and its component models, we can not generate paired samples for other models (i.e. ESA and NPY(n)). Instead, we follow Wang et al. (2010)'s method and first calculate the 95% confidence interval for

our proposed model. Then other systems can be compared with the joint model in this way: if the F-score of system **B** doesn't fall into the 95% confidence interval of system **A**, they are considered as statistically significantly different from each other.

For all significant tests, we measure the 95% confidence interval for the difference between two models. First, the test results show that "HDP+nVBE" and "HDP+HMM" are both significantly better than "HDP". Second, the "Joint" model significantly outperforms all its component models, including "HDP", "nVBE", "HDP+nVBE" and "HDP+HMM". Finally, the comparison also shows that the joint model significantly outperforms state-of-the-art systems like ESA and NPY(n).

5 Conclusion

In this paper, we proposed a joint model for unsupervised Chinese word segmentation. Our joint model is a combination of the HDP-based model, which is a word-based model, and HMM-based model, which is a character-based model. The way we combined these two component baselines makes it natural and simple to inference with Gibbs sampling. Then the joint model take advantage of a goodness-based method (nVBE) by using it to initialize the sampler. Experiment results conducted on PKU and MSRA datasets provided by the second SIGHAN Bakeoff show that the proposed joint model not only outperforms the baseline systems but also achieves better performance (F-Score) over several state-of-the-art systems. Significance tests showed that the improvement is statistically significant. Analysis also indicates that the joint model has a stronger ability to solve ambiguities in Chinese word segmentation. In summary, the joint model we proposed combines the strengths of character-based model, nonparametric Bayesian language model and goodness-based model.

Acknowledgments

The contact author of this paper, according to the meaning given to this role by Key Laboratory of Computational Linguistics, Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, is Baobao Chang. And this work is supported by National Natural Science Foundation of China under Grant No. 61273318 and National Key Basic Research Program of China 2014CB340504.

References

George Casella, Edward I. George. 1992. Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167-174.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 133. MLA.

Haodi Feng, Kang Chen, Xiaotie Deng, et al. 2004. Accessor variety criteria for Chinese word extraction *Computational Linguistics*, 30(1): 75-93.

Sharon Goldwater, Thomas L. Griffiths, Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1): 21-54.

Geoffrey E. Hinton. 1999. Products of experts. *Artificial Neural Networks*. Ninth International Conference on Vol. 1.

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

Zhihui Jin, Kumiko Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. *Proceedings of the COLING/ACL on Main conference poster sessions*, page 428-435.

Chunyu Kit, Yorick Wilks. 1999. Unsupervised learning of word boundary with description length gain. *Proceedings of the CoNLL99 ACL Workshop. Bergen, Norway: Association for Computational Linguis-tics*, page 1-6.

Pierre Magistry, Benoit Sagot. 2012. Unsupervized word segmentation: the case for mandarin chinese. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics*, page 383-387.

Daichi Mochihashi, Takeshi Yamada, Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. *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. Association for Computational Linguistics*, page 100-108.

Wenzhe Pei, Dongxu Han, Baobao Chang. 2013. A Refined HDP-Based Model for Unsupervised Chinese Word Segmentation. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer Berlin Heidelberg*, page 44-51.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, et al. 2006. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. *NIPS*.

Fuchun Peng, Fangfang Feng, Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. *Proceedings of COLING*, page 562-568.

Huihsin Tseng, Pichuan Chang, Galen Andrew, et al. 2005. A conditional random field word segmenter for sighan bakeoff 2005. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Vol. 171.

Kun Wang, Chengqing Zong, Keh-Yih Su. 2010. A character-based joint model for Chinese word segmentation. *Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics*, page 1173-1181.

Hanshi Wang, Jian Zhu, Shiping Tang, et al. 2011. A new unsupervised approach to word segmentation. *Computational Linguistics*, 37(3): 421-454.

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

Hai Zhao, Chunyu Kit. 2008. An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *IJCNLP*, page 6-16.

Ying Zhang, Stephan Vogel, Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? *LREC*.