A Refined HDP-Based Model for Unsupervised Chinese Word Segmentation

Wenzhe Pei, Dongxu Han, and Baobao Chang

Key Laboratory of Computational Linguistics, Ministry of Education,
Institute of Computational Linguistics, School of Electronics Engineering and
Computer Science, Peking University
{peiwenzhe, handx, chbb}@pku.edu.cn

Abstract. This paper proposes a refined Hierarchical Dirichlet Process (HDP) model for unsupervised Chinese word segmentation. This model gives a better estimation of the base measure in HDP by using a dictionary-based model. We also show that the initial segmentation state for HDP model plays a very important role in model performance. A better initial segmentation can lead to a better performance. We test our model on PKU and MSRA datasets provided by Second Segmentation Bake-off (SIGHAN 2005) [1] and our model outperforms the state-of-the-art systems.

1 Introduction

Chinese word segmentation is a very important component for almost all natural language processing tasks. Although supervised segmentation systems have been widely used, they rely on manually segmented corpora, which are often specific to domain and various kinds of segmentation guidelines. As a result, supervised segmentation systems perform poorly on out-of-domain corpus such as Microblog corpus [2] which contains lots of new words and domain specific words. In order to tackle this problem, unsupervised word segmentation becomes a very important issue. Various kinds of models have been proposed for unsupervised word segmentation task. [3] compared serveral popular models for unsupervised word segmentaion with a unified framework. [4] presented a model based on the Variation of Branching Entropy. [5] proposed a iterative model based on a new goodness algorithm that adopts a local maximum strategy and avoids thresholds.

In this paper, we present an unsupervised word segmentation method which refines the HDP-Based model [6]. This model gives a better estimation of the base measure in HDP by using a dictionary-based model. We also show that the initial segmentation state for HDP model plays a very important role in model performance. A better initial segmentation can lead to a better performance. We test our system on the PKU and MSRA benchmark datasets provided by Second Segmentation Bake-off (SIGHAN 2005) [1] and our method performed better than the state-of-the-art systems.

The remainder of this paper is structured as follows. In section 2, we give an overview of the HDP-based unsupervised word segmentation model. In section

3, we describe our models in detail. Section 4 shows our experiment results on the benchmark dataset. We then conclude the paper with section 5.

2 HDP-Based Unsupervised Word Segmentation

The Dirichlet Process (DP) is a stochastic process used in Bayesian non-parametric models of data. Let H be a distribution called base measure. The DP is a probability distribution, i.e. each draw from a DP is itself a distribution over distributions

$$G \sim DP(\alpha, H)$$

where H is basically the mean of the DP and α can be understood as an inverse variance. We can see that Dirichlet Process can be viewed as an infinite dimensional generalization of Dirichlet distributions.

The Hierarchical Dirichlet Process (HDP) is an extension to DP. It is a nonparametric Bayesian approach to clustering grouped data. It uses a Dirichlet process for each group of data, with the Dirichlet processes for all groups sharing a base distribution which is itself drawn from a Dirichlet process. The process defines a set of random probability measure G_j, one for each group, and a global random probability measure G_0. The global measure G_0 is distributed as a DP with concentration parameter α and base measure H and the random measure G_j are given by a DP with concentration parameter α_1 and base measure G_0

$$\begin{aligned} G_j &\sim DP(\alpha_1, G_0) \\ G_0 &\sim DP(\alpha, H) \end{aligned}$$

[6] proposed a Bayesian framework for unsupervised word segmentation with HDP. They define a bigram model by assuming each word has a different distribution over the words that follow it, but all these distributions are linked:

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

That is, $P(w_i | w_{i-1} = l)$ is distributed according to G_l which is a DP specific to word l. G_l is linked to other DPs by sharing a common base distribution G_0. The generating process can be represented according to the Chinese Restaurant Franchise (CRF) [7] metaphor. The metaphor is as follows. We have a restaurant franchise with a shared menu G_0 and each restaurant has infinitely many tables. When the $n + 1$th customer enter the restaurant l, the customer either joins an already occupied table k with probability proportional to the number n_{lk} of customers already sitting there and share the dish, or sits at a new table with probability proportional to α_1 and order a dish from menu G_0. Choosing dish from menu G_0 is a similar process, we can either choose an already ordered dish j with probability proportional to the number n_j of dishes already been ordered by all restaurants, or choose a new dish from H with probability proportional to α. In this bigram model, each w_{i-1} corresponds to a restaurant and each w_i is a dish. In practice, we can not observe G_l and G_0 directly because it will be infinite dimensional distribution over possible words. However, we can integrate out G_l and G_0 to get the posterior probability $P(w_i|w_{i-1} = l, h)$, where h is the observed segmentation result:

$$\begin{aligned} & P(w_i|w_{i-1} = l, h) \\ & = \int P(w_i|w_{i-1} = l, G_l)P(G_l|h)dG_l \\ & = \frac{n_{<w_{i-1},w_i>} + \alpha_1 P(w_i|h)}{n_l + \alpha_1} \end{aligned} \tag{1}$$

Here $n_{<w_{i-1},w_i>}$ is the number of occurrences of the bigram $< w_{i-1}, w_i >$ in the observed segment h and $P(w_i|h)$ is defined as:

$$\begin{aligned} & P(w_i|h) \\ & = \int P(w_i|G_0)P(G_0|h) \\ & = \frac{t_{w_i} + \alpha H(w_i)}{t + \alpha} \end{aligned} \tag{2}$$

Here t_{w_i} is the total number of tables labeled with w_i, t is the total number of tables and $H(w_i)$ is the prior knowledge of the probability of word w_i.

Given the equation of posterior probability, Gibbs sampling [8] is used for word segmentation by repeatly sampling the value of each possible word boundary location, conditioned on the current values of all other boundary locations. So each sample is from a set of two hypotheses: current location is either a word boundary or not. For example, let current segmentation result be $\beta c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}\gamma$, where β and γ are the sequence of words to the left and right of the area under consideration and $c_{i-2}c_{i-1}c_ic_{i+1}c_{i+2}$ forms a word w (Each c corresponds to a Chinese character). If the current sampling location i is a word boundary, the segmentation result would become $\beta w_1 w_2 \gamma$ where w_1=$c_{i-2}c_{i-1}c_i$ and w_2=$c_{i+1}c_{i+2}$. Otherwise, the segmentation result would remain the same. Let h_1 the first hypotheses and h_2 be the second. The posterior possibility for Gibbs sampling would be:

$$P(h_1|h^-) = P(w_1|w_l, h^-)P(w_2|w_1, h^-)P(w_r|w_2, h^-)$$

$$P(h_2|h^-) = P(w|w_l, h^-)P(w_r|w, h^-)$$

Here h^- is the current values of all other boundary locations without current position i and w_l (w_r) is the first word to the left (right) of the current word w. After the Gibbs sampling converged, the segmentation result can be obtained according to the word boundary results.

3 Refined Model

In this section, we present our model in detail and show how HDP-based model can be refined to impove the segmentation performance.

3.1 Improved Base Measure

As we can see in equation (1) and (2), the effect of posterior possibility of HDP is in fact a kind of smoothing. The bigram probability is smoothed by backing off to the unigram model and the unigram is smoothed by the base measure H, namely the prior probabilities over words. If the lexicon is finite, we can use a uniform prior $H(w) = \frac{1}{|V|}$. However, as every substring could be a word, the lexicon will be countbaly infinite. So building an accurate H is very important for word segmentation. [6] used a unigram character-based language model. [9] used a uniform distribution over characters dependent on word length with a Poission distribution.

In this paper, we use a dictionary-based model for estimating H. The intuition behind of our method is that given a large segmented corpus, a better estimation of the probability of a word can be obtained by using maximum likelihood estimation which is much more accurate than a simple character-based unigram model. However, in an unsupervised word segmentation task, we do not have a segmented corpus for probability estimation. To get the segmented corpus in an unsupervised way, we can use other unsupervised word segmentation system to segment the corpus. Although this could be inaccurate, substrings that are recognized as words would tend to have a high probability in the segmented corpus. To obtain a better estimation, we use different unsupervised word segmentation models to segment the corpus and merge the results together. Because different models give a different view of what a word is. The substring which is a real word tends to be recognized by all the models thus having a high probability. On the other hand, substring that is not a word tends to appear in none of the models.

As can be seen in Fig.1, we first use different unsupervised word segmentation systems to segment the training corpus. Then the words whose frequency is bigger than a threshold are selected from all the segmentation results and we merge the results to form a dictionary, that is, the frequency of the same word from different results are added up. Given the dictionary of words with their frequency, the base measure H is defined as follows:

$$
\begin{aligned}
H(w_i) &= \gamma P_{ml}(w_i) + (1-\gamma) P_{smooth}(w_i) \\
P_{ml}(w_i) &= \frac{C_{w_i}}{\sum_i^{|V|} C_{w_i}} \\
P_{smooth}(w_i) &= (1-p_s)^{|w_i|-1} p_s \prod_j p(c_{ij})
\end{aligned}
$$

Here, $P_{ml}(w_i)$ is the maximum likelihood estimation of the word probability from the dictionary and $P_{smooth}(w_i)$ is the base measure defined by [6]. As defined in

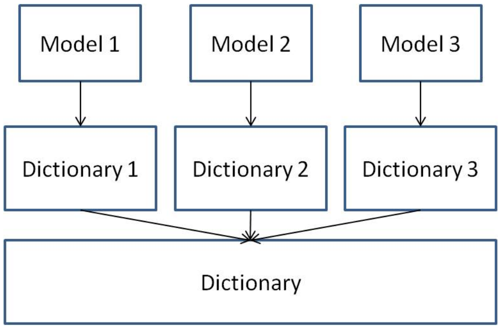

Fig. 1. Dictionary formed by exploiting different unsupervised word segmentation systems

[6], p_s is the probability of generating a word boundary. Thus $(1-p_s)^{|w_i|-1}p_s$ can be seen as the probability of $p(|w_i|)$ where $|w_i|$ is the length of word w_i. $p(c_{ij})$ is the probability of the jth character c_{ij} of word w_i, which can be obtained using maximum likelihood estimation from training data. $P_{ml}(w_i)$ and $P_{smooth}(w_i)$ are interpolated by parameter γ to make a trade-off between the two kinds of probability. As we can see in section 4, this better estimation of base measure H helps improve the model performance.

3.2 Initial State

As we present in section 2, the HDP model iteratively samples the value each possible word boundary location using Gibbs sampling. The procedure of sampling can be viewed as a random search in space of possible segmentation states. [6] use a random segmentation as the initial segmentation state and show that even with random initial state the model can still converge to a good result. However, we believe that a better initial state can lead to a better result and help converge much faster. In our method, we use a state-of-the-art unsupervised word segmentation system to segment the data first and use the segmentation result as a initial segmentation for the HDP model. As we can see in section 4, by using a better initial state, our method obtained a much better result than both the state-of-the-art system and the HDP model with random initial state.

4 Experiment

In this section we test our model on PKU and MSRA datasets released by the Second Segmentation Bake-off (SIGHAN 2005) [1] and make a comparision with previous work.

4.1 Prior Knowledge Used

Concerning unsupervised Chinese segmentation, a problem needs to be clarified is to what extent prior knowledge could be injected into the model. To be an as strict unsupervised model as possible, no prior knowledge such as word length, punctuation information, encoding scheme could be used. However, information like punctuation can be easily used to improve the performance. The problem is that we could not know what kind of prior knowledge other models used. For example, one might use manually designed regular expression to deal with numbers and dates, but does not list the regular expressions in paper. This makes it difficult to re-implement other models and make a fair comparision. To compare our model with previous models under the same condition, only puncuation information is used in our experiments. Punctuation information can improve the performance, since such information usually unambiguously marks boundary of words. It is very reasonable to use them in unsupervised Chinese segmentation model.

4.2 Model Selection

We randomly selected 2000 sentences from the training data as our development set for parameter tuning. We set α_1=100, α=10, γ=0.8 , p_s=0.5. We used two unsupervised word segmentation model to form the dictionary and give a initial segmentation result as described in section 3. The first model we use is nVBE [4]. It follows Harris's hypothesis in Kempe [10] and Tanaka-Ishii's [11] reformulation and base their work on the Variation of Branching Entropy. They improve on [12] by adding normalization and viterbi decoding. This model achieves state-of-the-art results on the Second Segmentation Bake-off (SIGHAN 2005) datasets. The second model we use is based on mutual information. Using mutual information is motivated by the observation of previous work by Hank and Church [13]. If character A and character B have a relatively high MI that is over a certain threshold, we prefer to identify AB as a word over those having lower MI values. We computed the mutual information on the training data. During the segmentation, we separate two adjacent characters to form a word boundary if their MI value is lower than a threshold. The threshold is set to 2.5 in our experiment. Although this model is not the state-of-the-art model, it is easy to implement and do give a different view of what a word is compared with nVBE. We put the training and test data together for segmenting. The word frequency threshold is set to 10 and two segmentations are merged to form the final dictionary.

4.3 Experiment Result

We test our model on the PKU and MSRA datasets released by the Second Segmentation Bake-off (SIGHAN 2005) [1]. We re-implement the nVBE model and the MI model and build our model based on these implementations. All the training data and test data are merged together for segmentation and only the test data are used for evaluation. The overall F-scores of different models are given in Table 1.

We can see that by using a dictionary-based model for estimating the base measure, the HDP model (HDP + dict) achieves a better result although only

Model	**PKU**	**MSRA**
HDP	68.7	69.9
nVBE[1]	77.9	78.2
ESA [5]	77.4	78.4
MI	66.1	70.2
HDP + dict	69.2	70.5
HDP + nVBE	79.2	79.4
HDP + MI	72.6	74.4
HDP + nVBE + dict	**79.3**	**79.8**

Table 1. Comparison of experiment results on PKU and MSRA datasets released by Second Segmentation Bake-off (SIGHAN 2005). ESA corresponds to the model in [5]

by a small margin. By using the segmentation result of nVBE as the initial segmentation, the HDP model (HDP+nVBE) gets a much better result than both the original HDP model and the nVBE model. Compared with nVBE, the F-score increases by 1.3% on PKU corpora and 1.2% on MSRA corpora. The HDP model with initial segmentation by MI (HDP+MI) also obtained a better result but not as well as HDP+nVBE model. This shows that the initial segmentation do play an important role in the model performance. A better initial segmentation tends to lead to a better performance. What's more, we find that with a better initial segmentation, the algorithm converges much faster than ordinary HDP. The HDP+nVBE converged after about 50 iterations while ordinary HDP needed 1000 iterations to converge. This saves a lot of time as sampling on a large dataset can be quite slow. The best model (HDP+nVBE+dict) is obtained by using the initial segmentation of nVBE and giving better estimation of base meausure with the dictionary-based model. Many errors are related to dates, Chinese numbers and English words. We believe that with a better preprocessing our model can achieve a much better result.

5 Conclusion

In this paper, we proposed a refined HDP model for unsupervised Chinese word segmentation. The refined HDP model uses a better estimation of base measure and replaces the random initial segmentation with a better one by exploiting other state-of-the-art unsupervised word segmentation systems. The refined HDP model achieves much better result than the state-of-the-art system on PKU and MSRA benchmark datasets.

[1] The results we got is slightly lower than the reported results in original paper. We have contacted the authors and they told us that the higher result they got was due to a bug in their code. Our results are considered to be reasonable with the bug free implementation

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No. 61273318 and 60975054

References

1. Emerson, T.: The second international chinese word segmentation bakeoff. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing. Volume 133. (2005)
2. Duan, H., Sui, Z., Tian, Y., Li, W.: The cips-sighan clp 2012 chinese word segmentation on microblog corpora bakeoff. (2012)
3. Zhao, H., Kit, C.: An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. In: The Third International Joint Conference on Natural Language Processing (IJCNLP-2008), Hyderabad, India. (2008)
4. Magistry, P., Sagot, B.: Unsupervized word segmentation: the case for mandarin chinese. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics (2012) 383–387
5. Wang, H., Zhu, J., Tang, S., Fan, X.: A new unsupervised approach to word segmentation. Computational Linguistics **37**(3) (2011) 421–454
6. Goldwater, S., Griffiths, T.L., Johnson, M.: A bayesian framework for word segmentation: Exploring the effects of context. Cognition **112**(1) (2009) 21–54
7. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association **101**(476) (2006)
8. Casella, G., George, E.I.: Explaining the gibbs sampler. The American Statistician **46**(3) (1992) 167–174
9. Xu, J., Gao, J., Toutanova, K., Ney, H.: Bayesian semi-supervised chinese word segmentation for statistical machine translation. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics (2008) 1017–1024
10. Kempe, A.: Experiments in unsupervised entropy-based corpus segmentation. In: Workshop of EACL in Computational Natural Language Learning. (1999) 7–13
11. Tanaka-Ishii, K.: Entropy as an indicator of context boundaries: An experiment using a web search engine. In: Natural Language Processing–IJCNLP 2005. Springer (2005) 93–105
12. Jin, Z., Tanaka-Ishii, K.: Unsupervised segmentation of chinese text by use of branching entropy. In: Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics (2006) 428–435
13. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational linguistics **16**(1) (1990) 22–29