Joint Learning of Chinese Words, Terms and Keywords

Ziqiang Cao[1] **Sujian Li**[1] **Heng Ji**[2]
[1]Key Laboratory of Computational Linguistics, Peking University, MOE, China
[2]Computer Science Department, Rensselaer Polytechnic Institute, USA
{ziqiangyeah, lisujian}@pku.edu.cn jih@rpi.edu

Abstract

Previous work often used a pipelined framework where Chinese word segmentation is followed by term extraction and keyword extraction. Such framework suffers from error propagation and is unable to leverage information in later modules for prior components. In this paper, we propose a four-level Dirichlet Process based model (DP-4) to jointly learn the word distributions from the corpus, domain and document levels simultaneously. Based on the DP-4 model, a sentence-wise Gibbs sampler is adopted to obtain proper segmentation results. Meanwhile, terms and keywords are acquired in the sampling process. Experimental results have shown the effectiveness of our method.

1 Introduction

For Chinese language which does not contain explicitly marked word boundaries, word segmentation (WS) is usually the first important step for many Natural Language Processing (NLP) tasks including term extraction (TE) and keyword extraction (KE). Generally, Chinese terms and keywords can be regarded as words which are representative of one domain or one document respectively. Previous work of TE and KE normally used the pipelined approaches which first conducted WS and then extracted important word sequences as terms or keywords.

It is obvious that the pipelined approaches are prone to suffer from error propagation and fail to leverage information for word segmentation from later stages. Here, we provide one example in the *disease* domain, to demonstrate the common problems in current pipelined approaches and propose the basic idea of our joint learning of words, terms and keywords.

Example: 血小板减少症(thrombocytopenia) 同(with) 类肝素(heparinoid) 有(have) 关系(relation).

This is a correctly segmented Chinese sentence. The document containing the example sentence mainly talks about the property of "类肝素(heparinoid)" which can be regarded as one keyword of the document. At the same time, the word 血小板减少症(thrombocytopenia) appears frequently in the *disease* domain and can be treated as a domain-specific term.

However, for such a simple sentence, current segmentation tools perform poorly. The segmentation result with the state-of-the-art Conditional Random Fields (CRFs) approach (Zhao et al., 2006) is as follows:

血小板(blood platelet) 减少(reduction) 症(symptom)
同类(of same kind) 肝(liver) 素有(always) 关系(relation)

where 血小板减少症 is segmented into three common Chinese words and 类肝素 is mixed with its neighbors.

In a text processing pipeline of WS, TE and KE, it is obvious that imprecise WS results will make the overall system performance unsatisfying. At the same time, we can hardly make use of domain-level and document-level information collected in TE and KE to promote the performance of WS. Thus, one question comes to our minds: can words, terms and keywords be jointly learned with consideration of all the information from the corpus, domain, and document levels?

Recently, the hierarchical Dirichlet process (HDP) model has been used as a smoothed bigram model to conduct word segmentation (Goldwater et al., 2006; Goldwater et al., 2009). Meanwhile, one strong point of the HDP based models is that they can model the diversity and commonality in multiple correlated corpora (Ren et al., 2008; Xu et al., 2008; Zhang et al., 2010; Li et al., 2012; Chang et al., 2014). Inspired by such existing work, we propose a four-level DP based model,

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1774–1778,
October 25-29, 2014, Doha, Qatar.

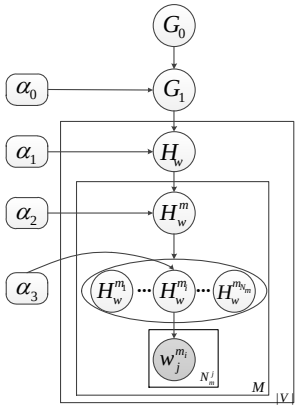

Figure 1: DP-4 Model

named DP-4, to adapt to three levels: corpus, domain and document. In our model, various DPs are designed to reflect the smoothed word distributions in the whole corpus, different domains and different documents. Same as the DP based segmentation models, our model can be easily used as a semi-supervised framework, through exerting on the corpus level the word distributions learned from the available segmentation results. Referring to the work of Mochihashi et al. (2009), we conduct word segmentation using a sentence-wise Gibbs sampler, which combines the Gibbs sampling techniques with the dynamic programming strategy. During the sampling process, the importance values of segmented words are measured in domains and documents respectively, and words, terms and keywords are jointly learned.

2 DP-4 Model

Goldwater et al. (2006) applied the HDP model on the word segmentation task. In essence, Goldwater's model can be viewed as a bigram language model with a unigram back-off. With the language model, word segmentation is implemented by a character-based Gibbs sampler which repeatedly samples the possible word boundary positions between two neighboring words, conditioned on the current values of all other words. However, Goldwater's model can be deemed as modeling the whole corpus only, and does not distinguish between domains and documents. To jointly learn the word information from the corpus, domain and document levels, we extend Goldwater's model by adding two levels (domain level and document level) of DPs, as illustrated in Figure 1.

2.1 Model Description

M DPs (H_w^m;$1 \leq m \leq M$) are designed specifically to word w to model the bigram distributions in each domain and these DPs share an overall base measure H_w, which is drawn from $DP(\alpha_0, G_1)$ and gives the bigram distribution for the whole corpus. Assuming the m^{th} domain includes N_m documents, we use $H_w^{m_j}$ ($1 \leq j \leq N_m$) to model the bigram distribution of the i^{th} document in the domain. Usually, given a domain, the bigram distributions of different documents are not conditionally independent and similar documents exhibit similar bigram distributions. Thus, the bigram distribution of one document is generated according to both the bigram distribution of the domain and the bigram distributions of other documents in the same domain. That is, $H_w^{m_j} \sim g(\alpha_3, H_w^m, H_w^{m-j})$ where H_w^{m-j} represents the bigram distributions of the documents in the m^{th} domain except the j^{th} document. Assuming the j^{th} document in the m^{th} domain contains N_m^j words, each word is drawn according to $H_w^{m_j}$. That is, $w_i^{m_j} \sim H_w^{m_j} (1 \leq i \leq N_m^j)$. Thus, our four-level DP model can be summarized formally as follows:

$$G_1 \sim DP(\alpha_0, G_0); H_w \sim DP(\alpha_1, G_1)$$
$$H_w^m \sim DP(\alpha_2, H_w); H_w^{m_j} \sim g\left(\alpha_3, H_w^m, H_w^{m-j}\right)$$
$$w_i^{m_j} | w_{i-1} = w \sim H_w^d$$

Here, we provide for our model the Chinese Restaurant Process (CRP) metaphor, which can create a partition of items into groups. In our model, the word type of the previous word w_{i-1} corresponds to a restaurant and the current word w_i corresponds to a customer. Each domain is analogous to a floor in a restaurant and a room denotes a document. Now, we can see that there are $|V|$ restaurants and each restaurant consists of M floors. The m^{th} floor contains N_m rooms and each room has an infinite number of tables with infinite seating capacity. Customers enter a specific room on a specific floor of one restaurant and seat themselves at a table with the label of a word type. Different from the standard HDP, each customer sits at an occupied table with probability proportional to both the numbers of customers already seated there and the numbers of customers with the same word type seated in the neighboring rooms, and at an unoccupied table with probability proportional to both the constant α_3 and the probability that the

customers with the same word type are seated on the same floor.

2.2 Model Inference

It is important to build an accurate G_0 which determines the prior word distribution $p_0(w)$. Similar to the work of Mochihashi et al. (2009), we consider the dependence between characters and calculate the prior distribution of a word w_i using the string frequency statistics (Krug, 1998):

$$p_0(w_i) = \frac{n_s(w_i)}{\sum n_s(.)} \tag{1}$$

where $n_s(w_i)$ counts the character string composed of w_i and the symbol "." represents any word in the vocabulary V.

Then, with the CRP metaphor, we can obtain the expected word unigram and bigram distributions on the corpus level according to G_1 and H_w:

$$p_1(w_i) = \frac{n(w_i) + \alpha_0 p_0(w_i)}{\sum n(.) + \alpha_0} \tag{2}$$

$$p_2(w_i|w_{i-1} = w) = \frac{n_w(w_i) + \alpha_1 p_1(w_i)}{\sum n_w(.) + \alpha_1} \tag{3}$$

where the subscript numbers indicate the corresponding DP levels. $n(w_i)$ denotes the number of w_i and $n_w(w_i)$ denotes the number of the bigram $< w, w_i >$ occurring in the corpus. Next, we can easily get the bigram distribution on the domain level by extending to the third DP.

$$p_3^m(w_i|w_{i-1} = w) = \frac{n_w^m(w_i) + \alpha_2 p_2(w_i|w_{i-1})}{\sum n_w^m(.) + \alpha_2} \tag{4}$$

where $n_w^m(w_i)$ is the number of the bigram $< w, w_i >$ occurring in the m^{th} domain.

To model the bigram distributions on the document level, it is beneficial to consider the influence of related documents in the same domain (Wan and Xiao, 2008). Here, we only consider the influence from the K most similar documents with a simple similarity metric $s(d_1, d_2)$ which calculates the Chinese character overlap ratio of two documents d_1 and d_2. Let d_m^j denote the j^{th} document in the m^{th} domain and $d_m^j[k](1 \le k \le K)$ the K most similar documents. d_m^j can be deemed to be "lengthened" by $d_m^j[k](1 \le k \le K)$. Therefore, we estimate the count of w_i in d_m^j as:

$$t_w^{d_m^j}(w_i) = n_w^{d_m^j}(w_i) + \sum_k s(d_m^j[k], d_m^j) n_w^{d_m^j[k]}(w_i) \tag{5}$$

where $n_w^{d_m^j[k]}(w_i)$ denotes the count of the bigram $< w, w_i >$ occurring in $d_m^j[k]$. Next, we model the bigram distribution in d_m^j as a DP with the base measure H_w^m :

$$p_4^{d_m^j}(w_i|w_{i-1} = w) = \frac{t_w^{d_m^j}(w_i) + \alpha_3 p_3^m(w_i|w_{i-1})}{\sum t_w^{d_m^j}(.) + \alpha_3} \tag{6}$$

With CRP, we can also easily estimate the unigram probabilities $p_3^m(w_i)$ and $p_4^{d_m^j}(w_i)$ respectively on the domain and document levels, through combining all the restaurants.

To measure whether a word is eligible to be a term, the score function $TH^m(\cdot)$ is defined as:

$$TH^m(w_i) = \frac{p_3^m(w_i)}{p_1(w_i)} \tag{7}$$

This equation is inspired by the work of Nazar (2011), which extracts terms with consideration of both the frequency in the domain corpus and the frequency in the general reference corpus. Similar to Eq. 7, we define the function $KH^{d_m^j}(\cdot)$ to judge whether w_i is an appropriate keyword.

$$KH^{d_m^j}(w_i) = \frac{p_4^{d_m^j}(w_i)}{p_1(w_i)} \tag{8}$$

During each sampling, we make use of Eqs. (7) and (8) to identify the most possible terms and keywords. Once a word is identified as a term or keyword, it will drop out of the sampling process in the following iterations. Its CRP explanation is that some customers (terms and keywords) find their proper tables and keep sitting there afterwards.

2.3 Sentence-wise Gibbs Sampler

The character-based Gibbs sampler for word segmentation (Goldwater et al., 2006) is extremely slow to converge, since there exists high correlation between neighboring words. Here, we introduce the sentence-wise Gibbs sampling technique as well as efficient dynamic programming strategy proposed by Mochihashi et al. (2009). The basic idea is that we randomly select a sentence in each sampling process and use the Viterbi algorithm (Viterbi, 1967) to find the optimal segmentation results according to the word distributions derived from other sentences. Different from Mochihashi's work, once terms or keywords are

identified, we do not consider them in the segmentation process. Due to space limitation, the algorithm is not detailed here and can be referred in (Mochihashi et al., 2009).

3 Experiment

3.1 Data and Setting

It is indeed difficult to find a standard evaluation corpus for our joint tasks, especially in different domains. As a result, we spent a lot of time to collect and annotate a new corpus [1] composed of ten domains (including *Physics*, *Computer*, *Agriculture*, *Sports*, *Disease*, *Environment*, *History*, *Art*, *Politics* and *Economy*) and each domain is composed of 200 documents. On average each document consists of about 4800 Chinese characters. For these 2000 documents, three annotators have manually checked the segmented words, terms and keywords as the gold standard results for evaluation. As we know, there exists a large amount of manually-checked segmented text for the general domain, which can be used as the training data for further segmentation. As with other nonparametric Bayesian models (Goldwater et al., 2006; Mochihashi et al., 2009), our DP-4 model can be easily amenable to semi-supervised learning by imposing the word distributions of the segmented text on the corpus level. The news texts provided by Peking University (named PKU corpus)[2] is used as the training data. This corpus contains about 1,870,000 Chinese characters and has been manually segmented into words.

In our experiments, the concentration coefficient (α_0) is finally set to 20 and the other three ($\alpha_{1\sim3}$) are set to 15. The parameter K which controls the number of similar documents is set to 3.

3.2 Performance Evaluation

The following baselines are implemented for comparison of segmentation results: (1) Forward maximum matching (FMM) algorithm with a vocabulary compiled from the PKU corpus; (2) Reverse maximum matching (RMM) algorithm with the compiled vocabulary; (3) Conditional Random Fields (CRFs)[3] based supervised algorithm trained from the PKU corpus; (4) HDP based semi-supervised algorithm (Goldwater et al., 2006) using the PKU corpus. The strength of Mochihashi et al. (2009)'s NPYLM based segmentation model is its speed due to the sentence-wise sampling technique, and its performance is similar to Goldwater et al. (2006)'s model. Thus, we do not consider the NPYLM based model for comparison here. Then, the segmentation results of FMM, RMM, CRF, and HDP methods are used respectively for further extracting terms and keywords. We use the mutual information to identify the candidate terms or keywords composed of more than two segmented words. As for DP-4, this recognition process has been done implicitly during sampling. To measure the candidate terms or keywords, we refer to the metric in Nazar (2011) to calculate their importance in some specific domain or document.

The metrics of F_1 and the out-of-vocabulary Recall (OOV-R) are used to evaluate the segmentation results, referring to the gold standard results. The second and third columns of Table 1 show the F_1 and OOV-R scores averaged on the 10 domains for all the compared methods. Our method significantly outperforms FMM, RMM and HDP according to t-test (p-value ≤ 0.05). From the segmentation results, we can see that the FMM and RMM methods are highly dependent on the compiled vocabulary and their identified OOV words are mainly the ones composed of a single Chinese character. The HDP method is heavily influenced by the segmented text, but it also exhibits the ability of learning new words. Our method only shows a slight advantage over the CRF approach. We check our segmentation results and find that the performance of the DP-4 model is depressed by the identified terms and keywords which may be composed of more than two words in the gold standard results, because the DP-4 model always treats the term or keyword as a single word. For example, in the gold standard, "岭南文化((Lingnan Culture)" is segmented into two words "岭南" and "文化", "数据接口(data interface)" is segmented into "数据" and "接口" and so on. In fact, our segmentation results correctly treat "岭南文化" and "数据接口" as words.

To evaluate the TE and KE performance, the top 50 (TE-50) and 100 (TE-100) accuracy are measured for the identified terms of one domain, while the top 5 (KE-5) and 10 (KE-10) accuracy for the keywords in one document, are shown in the right four columns of Table 1. We can see that DP-

[1] Nine domains are from http://www.datatang.com/data/44139 and we add an extra *Disease* domain.

[2] http://icl.pku.edu.cn

[3] We adopt CRF++(http://crfpp.googlecode.com/svn/trunk/doc/index.html)

4 performs significantly better than all the other methods in TE and KE results.

As for the ten domains, we find our approach behaves much better than the other approaches on the following three domains: *Disease*, *Physics* and *Computer*. It is because the language of these three domains is much different from that of the general domain (PKU corpus), while the rest domains are more similar to the general domain.

Method	F1	OOV-R	TE-50	TE-100	KE-5	KE-10
FMM	0.796	0.136	0.420	0.360	0.476	0.413
RMM	0.794	0.136	0.424	0.352	0.478	0.414
HDP	0.808	0.356	0.672	0.592	0.552	0.506
CRF	0.817	0.330	0.624	0.560	0.543	0.511
DP-4	**0.821**	**0.374**	**0.704**	**0.640**	**0.571**	**0.545**

Table 1: Comparison of WS, TE and KE Performance (averaged on the 10 domains).

4 Conclusion

This paper proposes a four-level DP based model to construct the word distributions from the corpus, domain and document levels simultaneously, through which Chinese words, terms and keywords can be learned jointly and effectively. In the future, we plan to explore how to combine more features such as part-of-speech tags into our model.

Acknowledgments

We thank the three anonymous reviewers for their helpful comments. This work was partially supported by National High Technology Research and Development Program of China (No. 2012AA011101), National Key Basic Research Program of China (No. 2014CB340504), National Natural Science Foundation of China (No. 61273278), and National Key Technology R&D Program (No: 2011BAH10B04-03). The contact author of this paper, according to the meaning given to this role by Peking University, is Sujian Li.

References

Baobao Chang, Wenzhe Pei, and Miaohong Chen. 2014. Inducing word sense with automatically learned hidden concepts. In *Proceedings of COLING 2014*, pages 355–364, Dublin, Ireland, August.

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.

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.

Manfred Krug. 1998. String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics*, 26(4):286–320.

Jiwei Li, Sujian Li, Xun Wang, Ye Tian, and Baobao Chang. 2012. Update summarization using a multi-level hierarchical dirichlet process model. In *Proceedings of Coling 2012*, pages 1603–1618, Mumbai, India.

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.

Rogelio Nazar. 2011. A statistical approach to term extraction. *IJES, International Journal of English Studies*, 11(2):159–182.

Lu Ren, David B. Dunson, and Lawrence Carin. 2008. The dynamic hierarchical dirichlet process. In *Proceedings of the 25th international conference on Machine learning*, pages 824–831.

Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269.

Xiaojun Wan and Jianguo Xiao. 2008. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860.

Tianbing Xu, Zhongfei Zhang, Philip S. Yu, and Bo Long. 2008. Dirichlet process based evolutionary clustering. In *ICDM'08*, pages 648–657.

Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. 2010. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1079–1088, New York, NY, USA.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 1082117.