Nonparametric Symmetric Correspondence Topic Models for Multilingual Text Analysis

Rui Cai(✉), Miaohong Chen, and Houfeng Wang

Key Laboratory of Computational Linguistics,
Peking University, Ministry of Education, Beijing, China
{cairui,wanghf}@pku.edu.cn, chenmiaohongnlp@gmail.com

Abstract. Topic model aims to analyze collection of documents and has been widely used in the fields of machine learning and natural language processing. Recently, researchers proposed some topic models for multilingual parallel or comparable documents. The symmetric correspondence Latent Dirichlet Allocation (SymCorrLDA) is one such model. Despite its advantages over some other existing multilingual topic models, this model is a classic Bayesian parametric model, thus can't overcome the shortcoming of Bayesian parametric models. For example, the number of topics must be specified in advance. Based on this intuition, we extend this model and propose a Bayesian nonparametric model (NPSymCorrLDA). Experiments on Chinese-English datasets extracted from Wikipedia (https://zh.wikipedia.org/) show significant improvement over SymCorrLDA.

Keywords: Multilingual text analysis · Topic model · Bayesian nonparametric model

1 Introduction

Getting valuable information from a large number of materials has attracted increasing research interest in recent years. A topic model is a type of statistical model for discovering the abstract topics that occur in a large collection of documents. In topic modeling, each document is regarded as a mixture of topics, and Latent Dirichlet Allocation (LDA)[2] is perhaps the most common topic model currently in use. Generally other models are extensions of LDA, most of which are appropriate for processing monolingual texts while some can be used for multilingual parallel or comparable documents to capture the statistical dependencies between multiple representations. Both parallel documents and comparable documents are merged documents consisting of multiple language parts, and parts of the former are translations from one language to another while the latter just describe similar concepts and events.

Bilingual topic models for bilingual parallel documents that have word-to-word alignments have been developed ([9]) and are directed towards machine translation. In contrast, some topic models focus on analyzing dependencies

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 270–281, 2015.
DOI: 10.1007/978-3-319-25207-0_23

among languages by modeling multilingual comparable documents, each of which consists of multiple language parts that are not translations of each other but instead describe similar concepts and issues. Conditionally Independent LDA (CI-LDA)[5] and SwitchLDA[6] (extension of CI-LDA) are multilingual topic models, but they only share per-document multinomial distributions between different languages, which produces weak dependencies. Correspondence LDA (CorrLDA)[7] and Symmetric Correspondence LDA (SymCorrLDA)[8] can also handle multilingual documents. In modeling, a pivot language plays a key role, text of which is translated to other languages, and the pivot selection is an important process to ensure estimation quality. SymCorrLDA incorporates a hidden variable to control the pivot language while CorrLDA must specify it in advance. In general, CorrLDA outperforms CI-LDA and SwitchLDA in processing comparable documents, and SymCorrLDA works more effectively than CorrLDA.

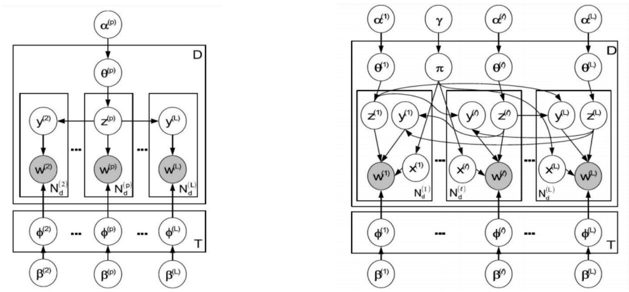
As a classic Bayesian parametric model, SymCorrLDA also bears the shortcomings of parametric models and the number of topics must be specified in advance. Nonparametric models can determine automatically the parameters scales and the complexity of models according to observed data while parametric models hold a strong assumption of data distribution. Hierarchical Dirichlet process (HDP) is a nonparametirc Bayesian model for clustering problems involving multiple groups of data. Our Nonparametric Symmetric Correspondence LDA (NPSymCorrLDA) is based on Hierarchical Dirichlet Process. Different from SymCorrLDA, NPSymCorrLDA generates a topic for each word according to the Dirichlet Process. The number of topics is open-ended, whose value grows at rate logarithmic in the number of words. Experiments on Chinese-English datasets extracted from Wikipedia show significant improvement over SymCorrLDA.

2 Multilingual Topic Models

Some researchers explored multilingual topic models that based on the premise of using multilingual dictionaries or WordNet ([10], [11], [12]). In contrast, CorrLDA and SymCorrLDA only require multilingual comparable documents that can be easily obtained. Below we introduce LDA-style topic models that handle multiple classes and can be applied to multilingual comparable documents.

2.1 Correspondence LDA(CorrLDA)

CorrLDA is a topic model for multilingual comparable documents proposed in [7]. This model first generates topics for one language part of a document. This language is referred as pivot language. For the other languages, CorrLDA uses the topics that are already generated in pivot language. Figure 1(a) shows a graphical model of CorrLDA. The p in the graph is the pivot language that is specified in advance. Algorithm 1 shows the process of generating a document according to CorrLDA model.



(a) (b)

Fig. 1. Graphical model representations of (a) CorrLDA,(b) SymCorrLDA.

Algorithm 1.. CorrLDA

for all D documents' pivot language part **do**
  Sample $\theta_d^p \sim \text{Dirichlet}(\alpha^p)$
end for
for all T topics and all L languages **do**
  Sample $\phi_t^l \sim \text{Dirichlet}(\beta^l)$
end for
for each of the N_d^p words w_i^p in language p of document d **do**
  Sample a topic $z_i^p \sim \text{Multinomial}(\theta_d^p)$
  Sample a word $w_i^p \sim \text{Multinomial}(\phi_{z_i^p}^p)$
end for
for each of the N_d^l words w_i^l in language l of document d **do**
  Sample a topic $y_i^l \sim \text{Uniform}(z_1^p, ..., z_{N_d^p}^p)$
  Sample a word $w_i^l \sim \text{Multinomial}(\phi_{y_i^l}^l)$
end for

2.2 Symmetric Correspondence LDA(SymCorrLDA)

Compared with CI-LDA and SwitchLDA, CorrLDA can capture more direct depency between languages. However as discussed before, it has to select the pivot language in advance. Since the pivot language may differ based on the subject, such as the country a document is about, it is often difficult to appropriately select the pivot language. To address this problem, Symmetric Correspondence LDA(SymCorrLDA) are proposed in [8]. Different from CorrLDA, SymCorrLDA can generates a flag that specifies a pivot language for each word, adjusting the probability of being pivot languages in each language part of a document according to a binomial distribution for bilingual data or a multinomial distribution for data of more than three languages. Figure 1(b) shows a graphical model of SymCorrLDA. Algorithm 2 shows the process of generating a document according to CorrLDA model. The pivot language flag x_i^l for an arbitrary language l indicates that the pivot language for the word w_i^l is its own language l, and

$x_i^l = m$ indicates that the pivot language for w_i^l is another language m different from its own language l. The indicator function δ takes the value 1 when the designated event occurs and 0 if otherwise.

Algorithm 2.. SymCorrLDA

for all D documents' pivot language part **do**
  for all L languages **do**
    Sample $\theta_d^l \sim$ Dirichlet(α^l)
  end for
  Sample $\pi_d \sim$ Dirichlet(γ)
end for
for all T topics and all L languages **do**
  Sample $\phi_t^l \sim$ Dirichlet(β^l)
end for
for each of the N_d^p words w_i^p in language p of document d **do**
  Sample a pivot language flag $x_i^l \sim$ Multinomial(π_d)
  if $x_i^l == l$ **then**
    Sample a topic $z_i^l \sim$ Multinomial(θ_d^l)
  else if $x_i^l == m$ **then**
    Sample a topic $y_i^l \sim$ Uniform$(z_1^m, ..., z_{M_d^m}^m)$
  end if
  Sample a word topic $w_i^l \sim$ Multinomial$(\delta_{x_i^l=l}\phi_{z_i^l}^l + (1 - \delta_{x_i^l=l})\phi_{y_i^l}^l)$
end for

3 Nonparametric Symmetric Correspondence LDA(NPSymCorrLDA)

When the SymCorrLDA model is applied to parallel or comparable documents, the number of topics(hyper parameter T) must be specified in advance. In practice, the number of topics often differs based on the scale of corpus. To find the best assignment of T, one natural method is to try different values and pick the best one. However, this is often time-consuming due to the high time complexity of model inference. Since SymCorrLDA uses Dirichlet distribution(DD) as the prior of parameters, it is a Bayesian parametric model. In this paper, we propose a nonparametric Bayesian model in the hope of overcoming the disadvantages mentioned above as well as achieving improvement over SymCorrLDA. Our model is based on Hierarchical Dirichlet Process(HDP)[13], which is a nonparametric Bayesian approach to cluster grouped data in statistics and machine learning. Here, we give a brief introduction of DP and HDP.

3.1 Dirichlet Process(DP)

The Dirichlet Process is a stochastic process used in probability theory. Let G_0 be a probability distribution and α be a positive real number. A Dirichlet Process

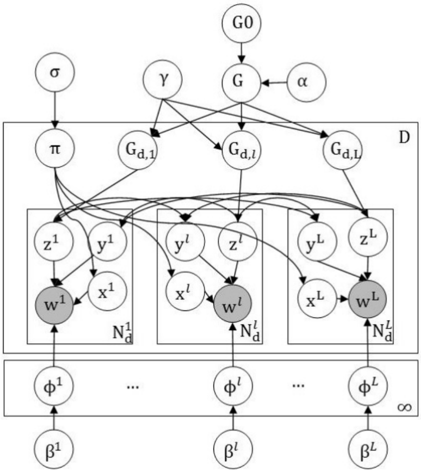


Fig. 2. Graphical model representations of NPSymCorrLDA.

G is defined as

$$G \sim DP(\gamma, G_0) \tag{1}$$

where γ is called the concentration parameter and G_0 the base measure. DP can be presented as the sticking-breaking process or the Chinese Restaurant Process(CRP). CRP is often used when we apply Gibbs sampling. In addition, Dirichlet Process is viewed as the infinite dimensional generalization of Dirichlet distribution. As with all Bayesian nonparametric models, DP's parameters grow as more data are observed. DP-based models have gained considerable popularity in the field of machine learning because of the above-mentioned flexibility([14,15]), especially in unsupervised learning. The flexibility also makes it an ideal candidate for clustering problems where the distinct number of clusters is unknown beforehand.

3.2 NPSymCorrLDA Based on HDP

The hierarchical Dirichlet process(HDP) is an extension to DP and often used to model grouped data. It uses Dirichlet process(DP) for each group of data, and all Dirichlet Process for grouped data share the same Dirichlet process as their base measure. For instance, for each group labeled j:

$$G_j \sim DP(\alpha, G) \tag{2}$$

where G itself is a DP

$$G \sim DP(\gamma, G_0) \tag{3}$$

Here α and γ are concentration parameters and H is a base measure. Like DP, HDP can be presented as sticking-breaking construction or Chinese Restaurant

Franchise(CRF)[13]. CRF is a generative process shown in Figure 3 and the metaphor is as follows: we have J restaurants, each with n_j customers(ϕ_{ji}'s), who sit at tables(ψ_{jt}'s). Now each table is served a dish (θ_k's) from a menu common to all restaurants. Suppose customer i sat at table t_{ji}. The conditional distributions according to $DP(\alpha, G)$ are:

$$t_{ji}|t_{j1}, ..., t_{ji-1}, \alpha \sim \sum_t \frac{n_{jt}}{\sum_{t'} n_{jt'} + \alpha}\delta_t + \frac{\alpha}{\sum_{t'} n_{jt'} + \alpha}\delta_{t^{new}} \tag{4}$$

where n_{jt} is the number of customers currently at table t. Let m_k denote the number of tables enjoying dish k. The customer i has to select a dish from the global dish menu if he chooses a new table, which is again distributed according to $DP(\gamma, G_0)$. A dish θ_k that has already been shared in the global menu would be chosen with probability $m_k/(\sum_{k'} m_{k'} + \gamma)$ and a new dish with probability $\gamma/(\sum_{k'} m_{k'} + \gamma)$.

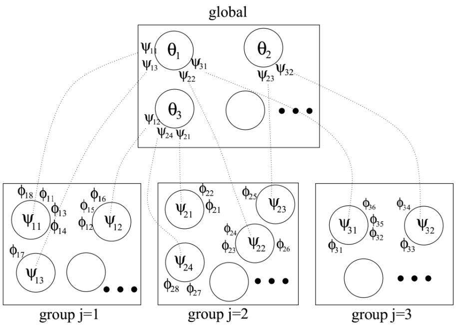


Fig. 3. An instantiation of CRF representation for the 3 group HDP. Each of the 3 restaurants has customers sitting around tables ,and each table is served a dish.

In our nonparameteric symmetric correspondence topic model(NPSymCorr LDA), the prior and posterior are not probability distributions any more, but stochastic process. This allows the model to determine the scale of parameters according to the data it observes. Like the SymCorrLDA, our model generates a flag that specifies a pivot language for each word, adjusting the probability of being pivot languages in each language part of a document according to a multinomial distribution over languages. This makes it possible for the model to estimates the best pivot language at the word level in each document. Unlike the SymCorrLDA, when the pivot flag is assigned with x, NPSymCorrLDA generates a topic for each word in document m according to the Dirichlet process Gm, x, and all the DPs share the same base measure G. Figure 2 shows a graphical model representation of NPSymCorrLDA. Algorithm 3 shows the process of generating a document according to NPSymCorrLDA model.

We use Gibbs sampling for model inference, and we sample from the posterior of HDF using CRF representation. The full conditional probability for

Algorithm 3.. NPSymCorrLDA

for all D documents' pivot language part **do**
  for all L languages **do**
    Sample $G_{d,l} \sim \mathrm{DP}(\gamma, G)$
  end for
  Sample $\pi_d \sim \mathrm{Dirichlet}(\sigma)$
end for
for all ∞ topics and all L languages **do**
  Sample $\phi_k^l \sim \mathrm{Dirichlet}(\beta^l)$
end for
for each of the N_d^l words w_i^l in language l of document d **do**
  Sample a pivot language flag $x_i^l \sim \mathrm{Multinomial}(\pi_d)$
  if $x_i^l == l$ **then**
    Sample a topic $z_i^l \sim G_{d,l}$
  else if $x_i^l == m$ **then**
    Sample a topic $y_i^l \sim \mathrm{Uniform}(z_1^m, ..., z_{M_d^m}^m)$
  end if
  Sample a word topic $w_i^l \sim \mathrm{Multinomial}(\delta_{x_i^l=l}\phi_{z_i^l}^l + (1-\delta_{x_i^l=l})\phi_{y_i^l}^l)$
end for

collapsed Gibbs sampling of our model is given by the equations below. The hyper-parameters are omitted for convenience.

$$P(z_i^l = k, x_i^l = l | w^l, x_{-i}, z_{-i}^l) \propto (n_{d,-i}^l + \sigma) \cdot P(z_i^l = k | z_{-i}^l) \cdot f_k^l(w_i^l) \tag{5}$$

$$P(y_i^l = k, x_i^l = m | w^l, x_{-i}, z^m) \propto (n_{d,-i}^l + \sigma) \cdot \frac{n_{d,k}^m}{\sum_{k=1}^T n_{d,k}^m} \cdot f_k^l(w_i^l) \tag{6}$$

where

$$f_k^l(w_i^l) = \frac{n_{k,-i}^l + \beta_k^l}{\sum_{k=1}^T n_{k,-i}^m l + \beta_k^l} \tag{7}$$

The conditional posterior probability $P(z_i^l = k | z_{-i}^l)$ is calculated according to the CRF representation of HDP.

4 Experiments

In this section, we demonstrate some examples with NPSymCorrLDA, and then we compare our model with baseline model using various evaluation methods. For evaluation, we use held-out log-likehood on English and Chinese dataset, the task of finding an English article that is on the same topic as that of a Chinese article, and a task with the languages reversed.

4.1 Settings

The dataset used in this work is a collection of Wikipedia articles: it is in English and Chinese, and articles in this collection are connected across languages via specific inter-language links. We extracted text content from original Wikipedia articles, removed link information, revision history information and other useless parts. We use Wikipedia Extractor[1] for this purpose. Chinese articles are edited by simplified and traditional type, we used Opencc[2] to convert traditional type to simplified type. For English articles, we removed over 1000 types of standard stop words. As for Chinese articles, we obtained words from Chinese sentences without stopping tags, using Chinese word segmentation system ICTCLAS[3]. The statistics of the datasets after preprocessing are shown in Table 1. We assumed each set of Wikipedia articles connected via inter-language links between two languages as a comparable document that consists of two language parts. To compute held-out log-likelihood, we randomly divided each of the training documents at the word level into 90% training set and 10% held-out set. To carry out the evaluation in the task of finding couterpart articles , we randomly divided the Wikipedia document collection at the document level into 90% training documents and 10% test documents.

Table 1. Summary of bilingual data.

	Chinese	**English**
No. of documents	5,029	5,029
No. of word types(vocab)	60,149	63,677
No. of word tokens	4,187,721	5,614,182

We estimated our model and SymCorrLDA as a baseline, using collapsed Gibbs sampling with the training set. In addition, we estimated a special implementation of NPSymCorrLDA, setting π_d in a sample way for comparison, where the pivot language flag for each word is randomly selected according to the proportion of the length of each language part. For all the models, we set symmetric Dirichlet hyperparameters $\alpha = 0.1$, $\beta = 0.01$ and $\gamma = 1$. We imposed the convergence condition of collapsed Gibbs sampling, such that the percentage change of held-out log-likelihood is less than 0.1%

4.2 Pivot Assignments

Figure 4 demonstrates how the frequency distribution of the pivot language-flag(binomial) parameter $\pi_{d,1}$ for the Chinese language with the bilingual dataset in NPSymCorrLDA changes while in iterations of collapsed Gibbs sampling. This figure shows that the piovt language flag is randomly assigned at the initial state, and then it converges to an appropriate bias for each document as the iterations proceed. We next demonstrate how the pivot language flags are assigned to each document. Figure 5 shows the title of eight documents and the corresponding $\pi_{d,1}$ when using the bilingual data. If $\pi_{d,1}$ is close to 1, the article can be consider

[1] http://medialab..di.unipi.it/wiki/
[2] https://code.google.com/p/opencc/
[3] http://ictclas.nlpir.org/

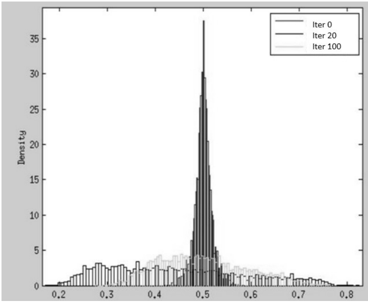


Fig. 4. Change of frequency distribution of $\pi_{d,1}$ according to number of iterations.

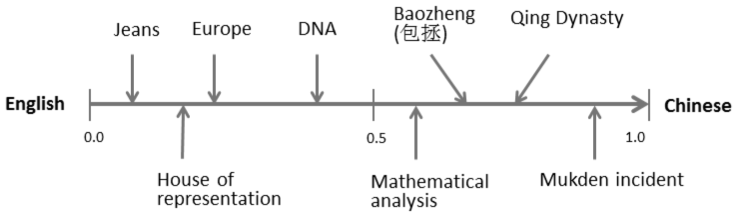


Fig. 5. Document titles and corresponding $\pi_{d,1}$

to be more related to a subject on Chinese or China. In contrast, if $\pi_{d,1}$ is close to 0 and therefore $\pi_{d,2} = 1 - \pi_{d,1}$ is close to 1, the article can be considered to be more related to a subject on English or English-speaking countries.

We further demonstrate the proportions of pivot assignments at the topic level. Figure 6 shows the content of 6 topics through 10 words with the highest probability for each language and for each topic when using the bilingual data, some of which are biased to Chinese or English while the others have almost no bias. It can be seen that the pivot bias to specific languages can be interpreted.

4.3 Held-Out Log-Likelihood

To estimate the performance of each topic model, we can calculate the held-out log-likelihood. The number of topics automatically converges to about 80 when training NPSymCorrLDA. We trained SymCorrLDA with setting topic numbers of 50 and 80 for comparsion. We know that the model with the higher held-out log-likelihood has greater predictive ability. The held-out set which we used in experiment is mentioned in Section 4.1.

The held-out log-likelihood of SymCorrLDA which shows superiority over other models is contained in [8]. We can see through Table 2 that our model NPSymCorrLDA is better than SymCorrLDA in both. Actually the improvements with NPSymCorrLDA are statistically significant compared to the previous model, according to the Wilcoxon signed-rank test at the 1% level in terms

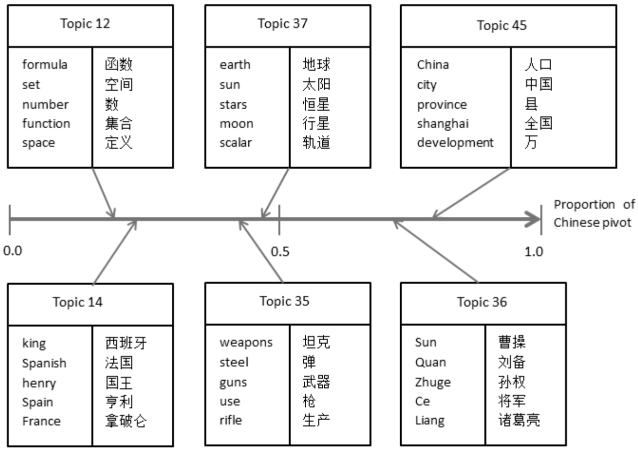


Fig. 6. Topic examples and corresponding proportion of pivots assigned to Chinese.

Table 2. Per-word held-out log-likelihood with bilingual data.

Methods	Chinese		English	
	T=50	T=80	T=50	T= 80
SymCorrLDA	-8.92	-8.83	-8.63	-8.55
NPSymCorrLDA	**-8.35**		**-8.27**	

of the word-by-word held-out log-likelihood. This indicates that our nonparametric model gets better estimation of the documents over SymCorrLDA. As for scalability, NPSymCorrLDA is as scalable as SymCorrLDA since the time complexity of these two models is the same.

4.4 Finding Counterpart Articles

Given an article, its hidden counterpart articles in other form of languages can be detected using a multilingual topic model. To fulfill this job, we conducted experiment with our bilingual dataset. We evaluated document-topic distribution of test documents for each language, using the topic-word distributions estimated by multilingual topic model with training documents. Using information mentioned above, we evaluated the performance of finding English counterpart articles using Chinese articles as queries, and vice versa. In our work, we utilized the technology of re-sampling to evaluate the document-topic distributions of test documents, and the process of re-sampling utilized the topic-word distribution estimated beforehand. Then we used the Jensen-Shannon(JS) divergence between a document-topic distribution of Chinese and that of English for each test document. JS divergence should be small if two counterpart articles have accurate latent topic evaluation. In light of this information, we first assumed each held-out Chinese article to be query and the corresponding English article

Table 3. MRR in Counterpart article finding task.

Methods	Chinese to English	English to Chinese
SymCorrLDA	0.3200	0.2948
NPSymCorrLDA	**0.3840**	**0.3814**

to be its counterpart, and then estimated the ranking of all the test articles of English in ascending order of the JS divergence, then we conducted the experiment with the languages reversed.

The mean reciprocal rank(MRR) are shown in Table 3. The definition of reciprocal rank is the multiplicative inverse of the rank of the counterpart article corresponding to each query article, and the MRR is the average of it over all the query articles. The topic number of SymCorrLDA is set to 80. And we can see through this table that our model shows its superiority over SymCorrLDA both with the Chinese and English queries. Statistically, the improvements with NPSymCorrLDA are significant according to the Wilcoxon signed-rank test at 1% level. Therefore, it is transparent that NPSymCorrLDA evaluates multilingual topics the most accurately in this experiment.

5 Conclusions

Many previous multilingual topic models connect assuming parallelism at the sentence level or word level, while works about document level is less than expected. SymCorrLDA incorporates a hidden variable to control a pivot language, in an extension of CorrLDA. However, SymCorrLDA has the same problem of how to assign the unknown parameters as other parameterized graph models. In this paper, we propose a nonparameterized symmetric correspondence LDA model(NPSymCorrLDA) that can be applied to multilingual documents, not using multilingual dictionaries. NPSymCorrLDA has an advantage that it does not require the number of topics specified in advance. We compared the performance of SymCorrLDA and NPSymCorrLDA in terms of held-out log-likelihood and in the task of cross-lingual link detection. Experiment results show that NPSymCorrLDA works significantly than SymCorrLDA. NPSymCorrLDA can be applied to other kinds of data that have multiple classes of representations, such as annotated image data. We plan to investigate this in future work.

Acknowledgments. Our work is supported by National High Technology Research and Development Program of China (863 Program) (No.2015AA015402), National Natural Science Foundation of China (No.61370117 & No.61433015) and Major National Social Science Fund of China (No.12&ZD227).

References

1. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM, 1999, August

2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of machine Learning research **3**, 993–1022 (2003)
3. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-vol. 2, pp. 880–889. Association for Computational Linguistics, 2009, August
4. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Mining multilingual topics from wikipedia. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1155–1156. ACM, April 2009
5. Stephen, E.E., Fienberg, S., Lafferty, J.: Mixed membership models of scientific publications. In: Proceedings of the National Academy of Sciences (2004)
6. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 680–686. ACM, August 2006
7. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 127–134. ACM, July 2003
8. Fukumasu, K., Eguchi, K., Xing, E.P.: Symmetric correspondence topic models for multilingual text analysis. In: Advances in Neural Information Processing Systems, pp. 1286–1294 (2012)
9. Zhao, B., Xing, E.P.: BiTAM: Bilingual topic admixture models for word alignment. In Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp. 969–976. Association for Computational Linguistics, July 2006
10. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 75–82. AUAI Press, June 2009
11. Jagarlamudi, J., Daumé III, H.: Extracting multilingual topics from unaligned comparable corpora. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 444–456. Springer, Heidelberg (2010)
12. Zhang, D., Mei, Q., Zhai, C.:. Cross-lingual latent topic extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1128–1137. Association for Computational Linguistics, July 2010
13. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the american statistical association **101**(476), (2006)
14. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics **9**(2), 249–265 (2000)
15. Rasmussen, C.E.: The infinite Gaussian mixture model. In: NIPS, vol. 12, pp. 554–560 (1999)