Khmer-Chinese bilingual LDA topic model based on dictionary

Xiaohui Liu and Xin Yan*

School of Information Engineering and Automation,
Kunming University of Science and Technology,
Kunming, Yunnan, 650500,China
Email: 623460648@qq.com
Email: kg_yanxin@sina.com
*Corresponding author

Guangyi Xu

Yunnan Nantian Electronics Information Co., Ltd.,
Kunming, Yunnan, 650050, China
Email: ntxgy@163.com

Zhengtao Yu and Guangshun Qin

School of Information Engineering and Automation,
Kunming University of Science and Technology,
Kunming, Yunnan, 650500, China
Email: ztyu@hotmail.com
Email: 289528150@qq.com

Abstract: Multilingual probabilistic topic models have been widely used in topic of mining area in multilingual documents, this paper proposes the Khmer-Chinese bilingual latent Dirichlet allocation (KCB-LDA) model based on the bilingual dictionary. With the bilingual attribute of entries in dictionary, this method first maps the words expressing same semantic meaning to the concept abstract layer, then group concepts into the same topic space. Finally, documents in different languages will share the same latent topics. The same topics can be represented in both Chinese and Khmer jointly when given a bilingual corpus by the introduction of the concept layer. The experimental results show that our topic modelling approach has better predictive power.

Keywords: multilingual probabilistic topic models; bilingual dictionary; Khmer-Chinese bilingual latent Dirichlet allocation; KCB-LDA; concept.

Reference to this paper should be made as follows: Liu, X., Yan, X., Xu, G., Yu, Z. and Qin, G. (2019) 'Khmer-Chinese bilingual LDA topic model based on dictionary', *Int. J. Computing Science and Mathematics*, Vol. 10, No. 6, pp.557–565.

Biographical notes: Xiaohui Liu received her Bachelor's degree in Telecommunication Engineering from the Nanjing University of Posts and Telecommunications in 2013. She is currently a Master student of School of Information Engineering and Automation, Kunming University of Science and Technology. Her current research interest is natural language processing.

Xin Yan received her Master's degree in Information and Electronic Science from the Yunnan University in 1992. She is currently an Associate Professor in the School of Information Engineering and Automation, Kunming University of Science and Technology. Her current research interests are topic models and sentiment analysis.

Guangyi Xu received his Master's degree in Information and Electronic Science from the Yunnan University in 1991. He is a Senior Engineer in Yunnan Nantian Electronics Information Co. His current research interests are natural language processing and data mining.

Zhengtao Yu received his PhD in Computer Application Technology from the Beijing Institute of Technology in 2005. He is a Professor in the School of Information Engineering and Automation, Kunming University of Science and Technology. His current research interests are machine translation and sentiment analysis.

Guangshun Qin received his Bachelor's degree in Computer Science from the Kunming University of Science and Technology in 2013. He is currently a Master student of School of Information Engineering and Automation, Kunming University of Science and Technology. His current research interest is natural language processing.

1 Introduction

Probabilistic latent topic models are mainly used in cross-lingual event clustering, cross-lingual document classification, cross-lingual information retrieval, cross-lingual semantic similarity of words and so on.

In monolingual LDA topic models, per-document topic distribution is usually drawn from a Dirichlet distribution. However, for cross-lingual corpus, related words in two different languages rarely appear together in a monolingual article. Therefore, monolingual model can not assign these related words in different languages to the same topic. Thus, the probabilities of related words in different languages could hardly be consistent. For example, when we use the languages of Chinese and Khmer, the words like '**ល្អល្អ**' (好的, good) will seldom co-occur in a monolingual text and a Dirichlet distribution cannot make P(好|Z)and P(**ល្អ**|Z) to be consistent. Such pairs of words cannot be representative words in a single coherent topic. There are many solutions to this: Blei and Lafferty (2006) proposed the method of using multivariate normal distribution as substitute of Dirichlet distribution which can produce relevant multivariate normal random variables. Cohen and Smith (2009) effectively used the method to simultaneously learn grammars in two languages from non-parallel multilingual data. However, the inference is complex without conjugate. Ni et al. (2009, 2011) proposed a cross-lingual ML-LDA model and implemented training topics with different linguistic versions of Wikipedia data and then used these universal-topics to do cross lingual text classification. Vulic et al. (2013) incorporated a bilingual topic model which was trained on comparable corpora such as Wikipedia articles for cross-language information retrieval. Zhu et al. (2013) proposed novel algorithms of document similarity which adopted the bilingual LDA model proposed by Ni et al. (2009) to build comparable

corpus. Mimno et al. (2009) proposed the PLTM model which is consistent to ML-LDA basically. All the above three models have restrictions on the corpus which need to be document-aligned. For unaligned texts, the MuTo model from Boyd-Graber and Blei (2009) operate with matching instead of words. A matching consists of pairs of words that link the source vocabulary to the target. One word can only have one semantic meaning in the whole dataset. Jagarlamudi and Daumé (2010) proposed JointLDA model, they also observe each topic as a mixture over these matching (concepts), where a word can be mapped to different concepts also means it will have different semantic meanings.

For the Khmer language processing is still in its infancy, and we cannot get the same large-scale comparable corpus, this paper proposes a solution by introducing electronic dictionary to bridge across language barriers, with the bilingual dictionaries of entries, words expressing the same meaning can be mapped to the concept layer and then grouped into the same topic space. Aligned documents are not necessarily required. Concepts are divided according to semantic which will have multi-words in target language. All the words of the same concept are subject to the same Dirichlet distribution. Therefore, this method takes advantage of the statistical properties of the texts which meet the cross-language semantic consistency between themes and concepts more effectively, it can predict the topic structure of both Khmer and Chinese bilingual documents.

2 KCB-LDA model

The levels of LDA generative process are as follows: document-topic-word. To construct the Khmer-Chinese bilingual latent Dirichlet allocation (KCB-LDA) model, concept is introduced to definite the topic distribution. Each topic is a mixture of concepts instead of words in a monolingual language. Compared to the three-level of monolingual LDA, KCB-LDA extend them to four-level as: document-topic-concept-word.

Given a non-parallel corpus, we assume all the documents are collections of concept units instead of bag-of-words. So similar documents share identical topic distribution regardless of their languages and the universal topics can be represented in Khmer and Chinese.

2.1 Introducing the Khmer-Chinese bilingual dictionary

2.1.1 Extracting the concepts and choosing the translations

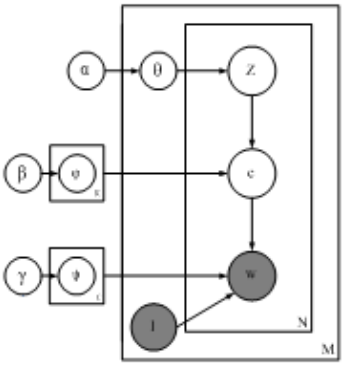
The concept layer targeted at word is language-independent, we introduce the terms from a bilingual dictionary in Khmer-Chinese. For the word in source language of Khmer, there are many corresponding translations of different meanings where exists polysemy. Concepts are extracted and grouped depending on different semantics, videlicet all entries are separated into different concepts by semantics. Thus, a word in source language belongs to many concepts is allowed. According to this principle, nearly 32000 concepts were collected from the bilingual dictionary.

A concept should consist of one word in source language and many related words in target languages in the model. They are always in the form as follows: ‘ការពារ: 防守, 防御, 保卫, 捍卫, 预防’(defend) from which we can achieve the per-concept translations distribution. Then a translation can be selected according to the chosen concept and the multinomial distribution of the corresponding translations.

2.1.2 Artificial addition to the Khmer-Chinese dictionary

Out-Of-Dictionary words cannot be generated by our model. If these words are ignored, consequently the topic distribution learned by our model is defective. We handle it by adding artificial dictionary entries. The artificial entries are monolingual and can only generate monolingual words. While the dictionary's are bilingual and can generate words in both source and target languages. All the artificial terms are represented in the form of 'w:_AR_' (_AR_:w), where 'AR' is the label of artificial entries. Under the excessive circumstances where the dictionary only has artificial terms, the one-to-many relation between words and concepts is reduced to one-to-one. The topic-concept distribution is weakened to the original topic-words distribution. Thus KCB-LDA is degenerated to monolingual LDA.

Figure 1 Graphical representation of KCB-LDA model



2.2 Constructing the KCB-LDA model

The detail description of KCB-LDA is as below (Figure 1):

1 For each document $d = 1 \ldots, M$, draw $\theta_m \sim Dir(\alpha)$. Randomly select a topic $z_i \sim \theta_m$ $Multinomial(\theta_m)$.

2 For each topic $z = 1 \ldots, K$, choose distribution over concepts $\varphi_k \sim Dir(\beta)$. Randomly select a concept c_i from the bilingual dictionary entries $c \sim Multinomial(\varphi_{z_i}) \cdot \xi(c_i, l_m)$.

3 For each concept $i = 1 \ldots, C$, choose distribution over translations $\psi_c \sim Dir(\gamma)$. Randomly select a word w_i according to $w \sim Multinomial(\psi_{c_i})$ when the language is l_T.

Where θ_m, φ_k, ψ_c represent document-topic, topic-concept and concept-word distribution respectively and are obtained by sampling from Dirichlet distribution. ($l_m \in \{S, T\}$), S and T represent source language and target language respectively. Concept distribution is

independent on language $c_i \sim Multinomial(\phi_{z_i})$. When generate a target language word, the model cannot choose the artificial dictionary entry like _AR_ : *w*, which can only generate source language word because word w does not have a corresponding target language translation. Therefore, the distribution should be $c_i \sim Multinomial(\varphi_{z_i}) \cdot \xi(c_i, l)$ where the function $\xi(c_i, l_m)$ is one, if concept c_i can generate a word from language l_m and zero otherwise. When language is target language l_T, w_i is generated according to the multinomial distribution over translations $w \sim Multinomial(\psi_c)$. When languages is l_S, the word can be directly determined, because a concept has only one word in the source language.

2.3 Inference

Our goal is getting the topic distribution of each document when given bilingual corpora. We use the collapsed Gibbs sampling to estimate the parameters (θ_m, φ_k). Once the topic of concept is fixed, (θ_m, φ_k) can be calculated by counting the concept frequency. The problem is transformed to calculate the conditional probability of topic sequence under the concept sequence then sample the topic sequence, the formula is as: $p(z \mid c) = \dfrac{p(c, z)}{\sum_z p(c, z)}$. The topic-concepts distribution of each token in every iteration is sampled from the probability as below:

$$p\left(Z_i = k, C_i = j \mid \mathbf{w}, \mathbf{z_i}, \mathbf{c_i}, \mathbf{l}\right) \propto \frac{n_{d,\neg i}^{(k)} + \alpha}{\sum_{z=1}^{K}\left(n_{d,\neg i}^{(z)} + \alpha_z\right)} \cdot \frac{n_{k,\neg i}^{(j)} + \beta}{\sum_{c=1}^{C}\left(n_{k,\neg i,}^{(c)} + \beta_c\right)} \cdot \frac{n_{c,\neg i}^{(h)} + \gamma}{\sum_{w=1}^{W}\left(n_{c,\neg i,}^{(w)} + \gamma_w\right)} \cdot p\left(w_i \middle| c_i, l_m\right) \tag{1}$$

where $n_{d,\neg i}^{(k)}$ denotes the number of the tokens that are assigned to topic $z_i = k$ in document d not counting the current word w_i, $n_{k,\neg i}^{(j)}$ is the number of concept $c_i = j$ that is associated with topic k in the entire corpus excluding the current, $n_{c,\neg i}^{(h)}$ is the times of word $w_i = h$ that is used along with concept c_i. $\sum_{z=1}^{K}\left(n_{d,\neg i}^{(z)} + \alpha_z\right)$ is the summation of $n_{d,\neg i}^{(z)}$ over all K topics, $\sum_{c=1}^{C}\left(n_{k,\neg i,}^{(c)} + \beta_c\right)$ counts the total number of concepts associated with topic k excluding the current, $\sum_{w=1}^{W}\left(n_{c,\neg i,}^{(w)} + \gamma_w\right)$ denotes the number of words in concept c_i in target language excluding the current, $p(w_i \mid c_i, l_m)$ is one when concept c_i can generate words of language l_m, else is zero.

When generate an unlisted word _AR_: w, the word is unique, the topic-concepts distribution of each token in every iteration is sampled from the probability as below:

$$p(Z_i = k, C_i = j \mid \mathbf{w}, \mathbf{z_i}, \mathbf{c_i}, \mathbf{l}) \propto \frac{n_{d,\neg i}^{(k)} + \alpha}{\sum_{z=1}^{K}\left(n_{d,\neg i}^{(z)} + \alpha_z\right)} \cdot \frac{n_{k,\neg i}^{(j)} + \beta}{\sum_{c=1}^{C}\left(n_{k,\neg i,}^{(c)} + \beta_c\right)} p\left(w_i \mid c_i, l_m\right) \tag{2}$$

3 Experiments and results

3.1 Experiment corpus

We crawled the Khmer and Chinese news corpus respectively from Cambodia daily website and Cambodian-news Chinese community website. The crawled Chinese news documents are largely related to the Khmer news documents from Cambodia daily. Thus we annotated document pairs which are topic related. We pre-processed the data by removing charts, tables, navigation, advertisements and other noise. Then the word segmentation and stop-words removal were performed on texts. Finally, we got nearly 50 M and total 5,024 aligned document-pairs which cover many fields, such as economy, politics, environment, sport, Internet and so on. The dictionary in this article is from the Khmer-Chinese dictionary published by Foreign Language Teaching and Research Press of China. Getting total 35,223 concepts by keyboard input, the dictionary coverage for experiments corpus vocabularies is about 70%.

3.2 Experimental design and results analysis

3.2.1 Model evaluation

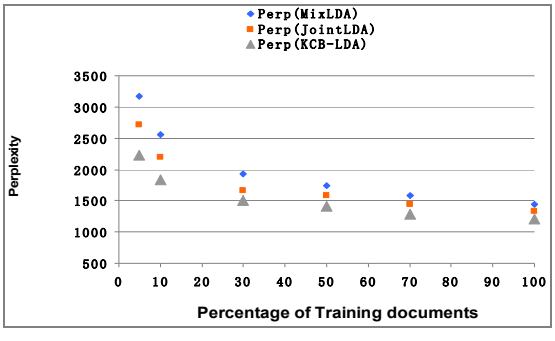
We use perplexity as metric to evaluate the predictive performance of generative models. Perplexity on the test sets can be calculated by formula (3). To verify the effectiveness of the proposed method, the experiments were conducted along with the control experiments of MixLDA a pseudo-multilingual probabilistic model and JointLDA in terms of perplexity scores. For KCB-LDA, $p(w_d) = p(w|l_d, d) = \sum_z p(w \mid k, l_d) p(k \mid d)$, where $p(w|k, l_d)$ is the sum of $p(c|k, l_d)$ which can generate the word w over all the dictionary entries. While for MixLDA which is a monolingual LDA trained on concatenated documents from aligned document pairs, we used the normal $p(w_d) = \sum_k p(w \mid k) p(k \mid d)$.

$$perp\left(D_{test}\right) = \exp\left(-\frac{\sum_d \log\left(p\left(w_d\right)\right)}{\sum_d N_d}\right) \tag{3}$$

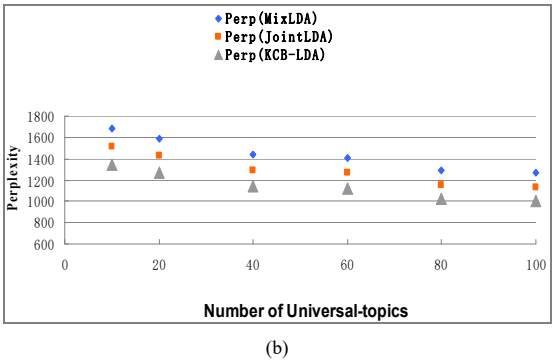
3.2.2 Parameters evaluation

In order to acquire the best conditions of the model, experiments are conducted about the topic number K, the percentage of test corpus and the iterations of Gibbs sampling. The experiments were conducted along with the control experiments of MixLDA model and JointLDA. The results are shown in Figures 2(a), 2(b) and 2(c).

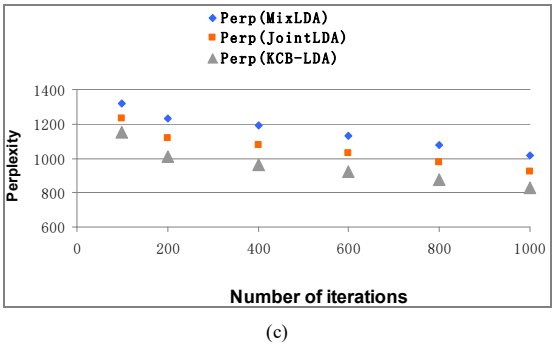
Figure 2 Perplexity scores comparison of MixLDA, JonitLDA and KCB-LDA, (a) perplexity vs. training document size (b) perplexity vs. topics (c) perplexity with iterations (see online version for colours)



(a)



(b)



(c)

In Figure 2(a), we vary the percentage of the corresponding training corpus. With the increasing percentage, perplexity scores become small, thus we fix the percentage at 100%.We show the Perplexity scatter diagram achieved by the topics from 10–120 to find the optimal number of topics in Figure 2(b). Experimental results show that as long as the number of selected topics is closed to the optimal number 100, topic extraction is generally satisfied. Figure 2(c) shows the perplexity scatter diagram achieved by the iterations from 100–1,000, as we can see the perplexity score is stabilised and achieve the best results when iterations are set to be 1,000. Figure 2 shows that compared with other models, our KCB-LDA model has better perplexity value, makes use of cross language corpora effectively and digs the document theme better.

3.2.3 Extracting multilingual topics

From the above experimental results we get the best and most appropriate conditions. The experiment was conducted according to these best conditions. The hyper parameters α, β and γ are set to be 50/K, 0.01 and 0.1 respectively, where K is the number of topics and is set to be 100. 70% of document tokens are randomly chosen for training. The number of iterations of the sampler is set 1,000 for convergence. Figure 3 shows some example universal-topics produced by KCB-LDA. The most likely words in each topic along with the translated words in the concepts are shown in Figure 3.

Figure 3 shows that KCB-LDA model can group the related words in different languages into the same topic: the corresponding entities ‘质检’ (quality inspection), ‘行业标准’ (industrial standard) which are bold do not exist in the Khmer-Chinese dictionary, words ‘管理’ (management), ‘检验’ (test) which are underlined are translations of the same word in source language while mean different semantics and belong to different concepts. It proves that our model can adapt to the situation of polysemy. The outputs are impressive and satisfy our expectation.

Figure 3 Topic-concept produced by KCB-LDA model

Topic 0	Topic 1	Topic 2	Topic 3
[កុងស៊ី]:[公司]	[ខុទ្ទកាល័យ]: [内阁，政府]	[កង្វក់]:[污染，弄脏]	[ការប្រកួតប្រជែង]:[比赛，竞赛]
[កុងត្រូល]:[管理，控制]	[គណបក្ស]:[党，党派，政党]	[កម្រើប]:[恶化]	[ចំណាត់ថ្នាក់]:[名次，排名]
[កម្មសិទ្ធិករ]:[所有者，物主，业主]	[កម្លាំង]:[权力，势力，威力，实力]	[ការ]:[保护，保卫，抵御]	[អនុវត្តន៍]:[训练]
[_AR_]:[质检]	[កាលៈទេសៈ]:[形势，情况，状况]	[កាលៈទេសៈ]:[环境]	[ពិន្ទុ]:[得分，比分]
[កើន]:[增加，增长]	[មានជម្លោះ]:[争议争端]	[កាយឫស្សទ្ធិ]:[健康]	[បញ្ជី]:[纪录]
[_AR_]:[行业标准]	*[ក]:[建设,建造,兴建,建立,创立]*	[ប្ដឹង]:[控诉，说]	[អ្នកអត្តពលកម្ម]:[运动员]
[កំប៉ាង]:[瑞尔，元]	[ការពារ]:[防守，防御，保卫，捍卫]	[កុងត្រូល]:[检验]	[កីឡាអូឡាំពិក]:[奥林匹克]
[ក្រឡា]:[面积，平方]	[កាប៊ីណេត]:[办公室，办公处，办公厅，事务所]	*[ក]:[建设,建造,兴建,建立,创立]*	[ការប្រកួត]:[锦标赛]
[កម្លង]:[集团]	[គមនាគមន៍]: [交往，交际]	[យានយន្ត]:[车辆]	[កីឡាហ្វូតបល់នៅអាមេរិចកាំ]:[足球]

4 Conclusions and future works

With the aligned information of entries in dictionary, our KCB-LDA model the topics on Khmer and Chinese documents and map the related words in both languages to common topics which achieve a cross-lingual topic model. Topics include the artificial entries which are not in the initial dictionary and thus achieve a cross-lingual semantic sharing. Experimental results show that the proposed method is suitable for discovering universal topics and obtain competitive results which can better predict unseen data. Of course the KCB-LDA model depends closely on the quality of dictionary which will affect the construction of the model. In the future of our work, we plan to do research on cross-lingual news event clustering according to the topic distribution of bilingual documents.

Acknowledgements

This work is supported by National Nature Science Foundation of China via grant 61462055, 61562049 , 61462054 and 61363044; New and High Technology Industry Development Project of Yunnan Province via grant 201606.

References

Blei, D.M. and Lafferty, J.D. (2006) 'Correlated topic models', in *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pp.147–154.

Boyd-Graber, J. and Blei, D.M. (2009) 'Multilingual topic models for unaligned text', in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pp.75–82.

Cohen, S.B. and Smith, N.A. (2009) 'Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction', in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp.74–82.

Jagarlamudi, J. and Daumé III, H.(2010) 'Extracting multilingual topics from unaligned comparable corpora', in *Proceedings of the 32th Annual European Conference on Advances in Information Retrieval*, pp.444–456.

Mimno D., Wallach, H.M., Naradowsky, J. et al. (2009) 'Polylingual topic models', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.880–889.

Ni, X., Sun, J-T., Hu, J. and Chen, Z. (2011) 'Cross lingual text classification by mining multilingual topics from Wikipedia', in *Proceedings of the 4th International Conference on Web Search and Web Data Mining*, pp.375–384.

Ni, X., Sun, J.T., Hu, J. et al. (2009) 'Mining multilingual topics from Wikipedia', in *Proceedings of the 18th International World Wide Web Conference*, pp.1155–1156.

Vulic, I., De Smet, W. and Moens, M.F. (2013) 'Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora', *Journal of Information Retrieval*, Vol. 16, No. 3,pp.331–368.

Wallach, H.M., Murray, I., Salakhutdinovn, R. et al.(2009) 'Evaluation methods for topic models', in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp.1105–1112.

Zhu, Z., Li, M., Chen, L. and Yang, Z. (2013) 'Building comparable corpora based on bilingual LDA model', *Proceedings of the ACL*, pp.278–282.