Online Chinese-Vietnamese Bilingual Topic Detection Based on RCRP Algorithm with Event Elements

Wen-xu Long[1,2], Ji-xun Gao[3], Zheng-tao Yu[1,2,*],
Sheng-xiang Gao[1,2], and Xu-dong Hong[1,2]

[1] School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650051, China
[2] The Intelligent Information Processing Key Laboratory,
Kunming University of Science and Technology, Kunming 650051, China
[3] School of Computer Science, Henan Institute of Engineering, Zhengzhou, 451191, China

Abstract. On account of the characteristics of online Chinese-Vietnamese topic detection, we propose a Chinese-Vietnamese bilingual topic model based on the Recurrent Chinese Restaurant Process and integrated with event elements. First, the event elements, including the characters, the place and the time, will be extracted from the new dynamic bilingual news texts. Then the word pairs are tagged and aligned from the bilingual news and comments. Both the event elements and the aligned words are integrated into RCRP algorithm to construct the proposed bilingual topic detection model. Finally, we use the model to determine if the new documents will be grouped into a new category or classified into the existing categories, as a result, to detect a topic. Through the contrast experiment, the proposed model achieves a good effect on topic detection.

Keywords: Topic model, Event elements, Storyline. Bilingual, RCRP.

1 Introduction

Vietnam is closely connected with China so that to timely and accurately detect Chinese-Vietnamese bilingual topic trend is of great significance to enhance the communication and cooperation between both sides. On monolingual topic detection,a large number of research has been done at home and abroad. Considering of the elements of news,Wang extracted the name entities and integrated them into LDA model to track a series of related news[1]. On bilingual topic detection, De Smet W. proposed an intermediate LDA model of English and Dutch,which were trained from English-Dutch word pairs of Wikipedia[2].Ni,et al,proposed a cross language classification model by minging multilingual topics from Wikipedia page and data[3].Considering of the dynamic topic detection,Ahmed integerated RCRP algorithm with LDA model according to the temporal of the news,which gained good effect[4-6].

Online Chinese-Vietnamese topic detection is to analysis the dynamic growing bilingual news text.Hence According to the characteristics of news,we need to combine the key elements of an event of who,when and where and analysis the text relevance by the constructed entities,eg.who,when and where;Also it should timely acquire the

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 422–429, 2014.

growing mixed bilingual news data and be able to analysis the data dynamically;Bilingual news has the characteristic of cross-language and we need reduce the error from direct translation. The core of the RCRP algorithm is a nonparametric Bayesian method,using the prior parameters of this time to estimate parameters of next period,from which can provide a dynamic anlysis in constant periods.RCRP has been a periodic and nonparametric evolving clustering method, which can file a new document into the existing cluster according to a prior probability. Meanwhile it is featured with the unfixed number of clusters and can get a cluster at any time based on a specific probability,which accord with the characteristic of the topic of randomly generated and developing,extincting with time[5-6]. Hence that according to the characteristics of online bilingual topic detection and the advantage of RCRP,we try to exploit a model based on RCRP algorithm,which integrates the temporal information, the entities and bilingual aligned words to solve the topic detection problem.

2 Online Bilingual LDA Integrated with Event Elements

2.1 To Integrate Time Series with RCRP

As a special case in the Dirichlet process, the recurrent Chinese restaurant process is on the basis of the LDA model[5-6] and according to the Markov Assumption, assume the parameters $\alpha_t|\alpha_{t-1}\sim N(\alpha_{t-1},\delta^2 I)$ and $\beta_t|\beta_{t-1}\sim N(\beta_{t-1},\delta^2 I)$,instead of assuming that α and β would remain unchanged at any point of time within a certain time frame.

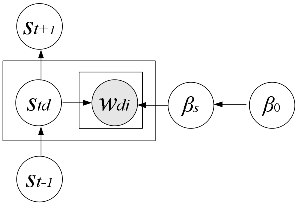

Fig. 1. RCRP algorithm graghical model

For time $t\in\{1,2,\dots,T\}$

(a) Get s_{td} by $s_{td}|s_{1:t-1},s_{t,1:d-1}$

(b) If s_{td} is a new storyline get $\beta_s|\beta_0$

(c) Get $w_{di}\sim\beta_0$ for news text d_i

2.2 To Integrate Event Elements

After an event has happened, various reports about this event will be made from every aspect. Although the words used and the opinions expressed in each report are different, a consensus is always reached on the key questions connected to the event. In fact, all of the event elements such as the time and entities including the name of the person, the name of the place and the name of the organization have become an important approach for the differentiation of news topics with different plots,which will be extracted from the news text.

In LDA model[7], the topic distribution in a document is subject to the wording condition and obtained through the calculation of the word frequency. But some words are useless or even bring deviation for topic detection.Therefore it's necessary to use the entities as the label information for the detection work.

2.3 To Integrate Aligned Bilingual Event Arguments

Huge work has been done on the fields of bilingual entity translation[8],word alignment[9] and cross-language entity linking[10].With the methods of these work and through the Chinese-Vietnamese page from Wikipedia, it's possible to get the comparable corpus. For word-to-word translation from Chinese V_s to Vietnamese V_t with m_D defined as (v_s, v_t) and $v_s \in V_s, v_t \in V_t$. And it's applicable to utilize the ontology-based event knowledge base to calculate the semantic similarity between the nouns, the verbs and the entities in the bilingual documents, selecting those pairs with high similarity as the collection of synonym pairs, m_K. Meanwhile with the application of the alignment method for the bilingual event elements contained in the news, the aligned elements will be obtained from the bilingual pages to constitute the collection of aligned bilingual event arguments, m_E.The total collection $m = m_D \cup m_K \cup m_E$,which is inputed into LDA to construct a bilingual topic model.We train Chinese and Vietnamsese data separately to get each posterior parameter γ and π,and get a joint distribution by the collection m.

2.4 The Proposed Online Bilingual LDA Integrated with Event Elements

As shown in Fig.2., it is probabilistic graphical model constructed for the Chinese-Vietnamese topics.

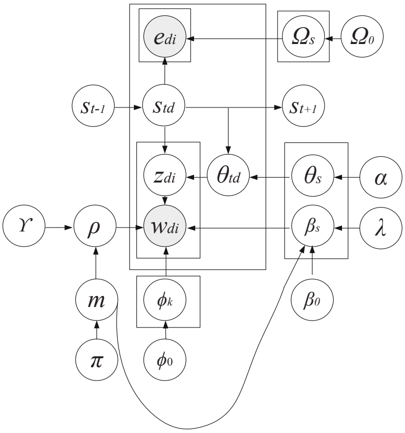

Fig. 2. Online bilingual LDA integrated with event elements

We use Gibbs Sampling to train the model using the joint distribution of bilingual information.The joint distribution is gotten from the algorithm below.

(a) Setting the topic number is *K*;
(b) Setting the parameter α_0 and ϕ_0 ;
 Sampling K times $\gamma \sim Dirichlet(\phi)$
 Sampling K times $\pi \sim Dirichlet(\phi)$
(3) For each $d \in \{d_1, \ldots, d_t\}$
 Sampling $\rho \sim Dirichlet(\alpha)$
(4) Draw the story indicator:
$s_{td}|s_{1:t-1}, s_{t,1:d-1} \sim RCRP(\gamma, \lambda, \Delta)$
(5) If s_{td} is a new story,
 Sampling K times over aligned words $\beta_{snew} \sim Dirichlet(\beta_0)$
 Sampling K times over entities $\Omega_{snew} \sim Dirichlet(\Omega_0)$
 Sampling K times over entities $\theta_{snew} \sim Dirichlet(\alpha)$
 Topic proportions $\theta_{td} \sim Dirichlet(\theta_{std})$
 Entities $e_{td} \sim Multinamial(\Omega_{std})$

Through the calculation of the posterior probability, $P(z_{1:T}, s_{1:T}|x_{1:T})$, we shall manage to realize the Chinese-Vietnamese topic detection with z_t, s_t, x_t representing separately the subject marker, the topic marker and the bag of words, which consists of the event elements w_{td}^{K+1} that cover the aligned bilingual entity e_{td} and the time etc. in the time slot of t. In a new document corresponding to the event s, the probability based on which x_t will be assigned with the ith topic among the existing K-1 topics within the time of t is shown as below:

$$P(z_{td} = z_{tdi}\big|s_{ts}^{-td}, x_{td}, rest) = \frac{C_{tdi}^{-i} + \frac{C_{si}^{-i}+\alpha}{C_{s.}^{-i}+\alpha(K+1)}}{C_{td.}^{-i}+1} \frac{C_{ix}+\phi_0}{C_{i.}+\phi_0 W} \tag{1}$$

Herein, C_{tdi}^{-i} refers to the number of the topics without the ith topic in the document d at the time, t. C_{sk}^{-i} represents the number of the topics corresponding to the event s without the ith topic, while C_{kxt}^{-i} stands for the number of the topics covered with the word x without the ith topic.

$$Ctd.–i=Ctdi–i,s.–i=Csi–i,Ci.=Cix \tag{2}$$

For the model we propose,

$$P(s_{td}\big|s_{t-\Delta:t}^{-td}, x_{td}, rest) =$$
$$P(s_{td}\big|s_{t-\Delta:t}^{-td})P(s_{td}\big|s_{t-\Delta:t}^{-td}, rest)P(z_{td}|s_{td}, rest)P(e_{td}|s_{td}, rest)P(w_{td}^{K+1}|s_{td}, rest) \tag{3}$$

Where rest denotes all other hidden variables not including z_t, s_t, x_t.

The posterior probability, P is computed using the chain rule as follows:

$$P(z_{1:T}, s_{1:T}|x_{1:T}) = \prod_{i=1}^{n_{td}} P\left(z_{tdi}\middle|s_{td} = s, z_{td}^{-td}, rest\right) \quad (4)$$

3 Experiments and Analysises

3.1 Evaluation Index

Regarding the clustering performance, generally tests have been conducted and the detection error cost, C_{det} has been applied to evaluate the effect and performance of the algorithm. Consisting of the miss rate P_{miss} of the model and the false detecting rate, P_{fa}, C_{det} has been considered as the evaluation criteria published by the National Institute of Standards and Technology (NIST) for the topics and tasks with the specific calculations described as below.

$$C_{det} = C_{miss} \times P_{miss} \times P_{target} + C_{fa} \times P_{fa} \times P_{non-target} \quad (5)$$

Herein, C_{miss} represents the cost coefficient of the missing detection with C_{fa} standing for the cost coefficient of the false detection, while P_{target} means the prior probability for the system to make positive judgment with $P_{non-target}$ representing the prior probability for the system to make negative judgment. However according to the standard made by NIST, we generally assume that C_{miss} and C_{fa} are separately 1 and 0.1 with the values for P_{target} and $P_{non-target}$ respectively being 0.02 and 0.98. Below please find the calculation formula of P_{miss} and P_{fa} with the parameters defined in the table as below.

During the application, generally evaluation is made according to the normalized detection error cost with the calculation formula given as below:

$$Norm(C_{det}) = \frac{C_{det}}{\min\left(C_{miss} \times P_{target}, C_{fa} \times P_{non-target}\right)} \quad (6)$$

3.2 Results and Analysises

In the experiment, 376274 news and reports have been fetched from the Chinese website and 221035 reports have been grabbed from the Vietnamese website to act as the data set. Half of the data and Wikipedia data is used to train the model.According to the web page tags, we could find that the topic, the release time and the content of the reports have been covered completely in the data acquired for the news. In fact, the experiment has been conducted for the following purposes on the basis of the dataset: (1)Observe the detection error cost for the model when the number of topics is different; (2) Observe the detection error cost when separately removing the event element, the time sequence and the bilingual word pairs from the model; (3) In the case that the number of the topics have been specified, make a comparison with the K-means clustering algorithm and the LDA model that hasn't been integrated with the features. For the hyper-parameters of the model, make sure that $\beta_0 = 0.1$, $\emptyset_0$=0.01, Ω_0=0.001, $\alpha_0 = \frac{0.1}{K+1}$, $\lambda = 0.5$ and $\Delta = 3$.

- Comparison of different assumpt topic numbers

In the experiment, we've also tested the effect of the model in the case of different number of topics K by assuming that the number of Gibbs sampling is N=500.

Table 1. The evaluation of the model in the case of different topic number

K	1	30	50	100	200	300
$\boldsymbol{C_{det}}$	0.79	0.78	0.740	0.697	0.733	0.738

Judging from the table as above, we could find that when K=100, the value of C_{det} is minimum and slow changes of the value of C_{det}have been found when *K*<50 or *K*>100. All of these prove that in a given dataset, there should be an appropriate *K* that might lead to the minimum C_{det} in the model. In fact, C_{det} of the model will be reduced with the increase in the number of the topics within a certain range. However when the number of the topics reaches a threshold, the influence on the model brought by the increase in the number will become weaker and weaker.

- With different features(event elements、storyline and bilingual word pairs)

In order to illustrate the role and importance of the three model features which have been integrated : the event elements, the time series and the collection of bilingual word pairs, we'd like to remove a certain element from the model for the contrast experiment in the test for the purpose to test the influence of each element on the model. Meanwhile in the test, we've determined that the number of the topic, K=100 and the sampling number N=500.

Table 2. The evaluation of the model in the case of integrating different features

Features	C_{det}
Storyline and bilingual word pairs	0.91
Event elements and bilingual word pairs	0.733
Event elements and Storyline	0.79

According to the result, we could find that the event elements have played a critical role in the topic detection. However the collection of bilingual word pairs has contributed a lot to the improvement of the model effect due to the reduction in the error caused by the polysemy when various translation tools are utilized.

- Compared with K-means clustering and simple LDA

In the test, first on account of the 1000 Chinese and Vietnamese news about the anti-Chinese movements in Vietnam and the other kinds of bilingual news with the amount of news up to 1000, we shall evaluate the performance of these three models regarding the anti-Chinese movements in Vietnam. During the application of K-means algorithm, we assume that the clustering number, K is set as 20, 30 and 50. While for the LDA model, we've assumed that the number of the topics is 100 with the sampling number at 300.

Table 3. The evaluation of diffent algorithm on topic detection

Algorithm		C_{det}
K-means	K=20	0.861
	K=30	0.847
	K=50	0.851
Simple LDA		0.89
Proposed model		0.714

Through the aforesaid tests, we could find that the online bilingual LDA model integrated with the event elements has been superior to the other two kinds of models/ algorithms from the perspective of the evaluation result. Limited by the computation process, the K-means algorithm is not applicable to massive dataset, the unknown clustering center and the changes in the incremental data. However if the LDA hasn't been integrated with the features, it means that the event elements haven't been covered in the model to bring noises to numerous words in the bag of words.

3.3 Experimental Evaluation

According to the test result, it turns out that: the event elements and the time sequence contained in the news and reports and the creation of bilingual words have influenced a lot the effect of the topic detection. The proposed model, which has been developed on the basis of the traditional LDA model and integrated with the event elements, the time series and the Chinese-Vietnamese word pairs, is able to cluster more accurately the data of the bilingual news.

4 Conclusion

In this paper, we propose a RCRP-based online Chinese-Vietnamese topic detection model according to the characteristics of dynamic bilingual news,which effectively integrates time series,event elements and bilingual information into one LDA model and achieves good effect in the experiment.Our work next step is to exploit a more advanced topic model by using bilingual machine translation and bilingual knowledge resources to calculate the relativity of news on text level to implement the news clustering.

Acknowledgments. This paper is supported by National Nature Science Foundation (No.61472168,61262041,61175068), and the key project of Yunnan Nature Science Foundation (2013FA030),and Science and technology innovation talents fund projects of Ministry of Science and Technology(No.2014HE001) .

References

1. Wang, D., Liu, W., Xu, W.: Topic Tracking Based on Event Network. In: 2011 4th International Conference on Cyber, Physical and Social Computing Internet of Things (iThings/CPSCom), pp. 488–493 (2011)
2. De Smet, W., Moens, M.F.: Cross-language linking of news stories on the web using interlingual topic modelling. In: Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, pp. 57–64. ACM (2009)
3. Ni., X., Sun, J.-T., Hu, J., Chen, Z.: Cross Lingual Text Classification by Mining Multilingual Topics From Wikipedia. In: Proceedings of the Fourth ACM International Confernce on Web Search and Data Mining, pp. 375–384. ACM (2011)
4. Ahmed, A., Xing, E.P.: Dynamic Non-parametric Mixture Models and the Recurrent Chinese Restaurant Process: With Applications to Evolutionary Clustering. In: SDM (2008)
5. Ahmed, A., Ho, Q., Eisenstein, J., et al.: Unified analysis of streaming news. In: Proceedings of the 20th International Conference on World Wide Web, pp. 267–276. ACM (2011)
6. Ahmed, Q., Ho, C., Teo, J., Eisenstein, A.J., Smola, E.P.: Xing The Online Infinite Topic-Cluster Model: Storylines From Streaming Text. CMU-ML-11-100 (2011)
7. Blei, D.M., Andrew, Y.N., Michael, I.J.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
8. Sproat, R., Tao, T., Zhai, C.X.: Named Entity Transliteration with Comparable Corpora. In: Proceeding ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 73–80 (2006)
9. Espla-Gomis, M., Sanchez-Martinez, F., Forcada, M.L.: A Simple Approach to Use Bilingual Information Sources for Word Alignment. Procesamiento del Lenguaje Natural, 93–100 (2012)
10. Fahrni, A., Strube, M.: HITS' Cross-lingual Entity Linking System at TAC 2011:One Model for All Languages. In: Proceeding of Text Analysis Conference, November 14-15 (2011)