Detecting Common Discussion Topics Across Culture From News Reader Comments

Bei Shi[1], Wai Lam[1], Lidong Bing[2] and Yinqing Xu[1]
[1]Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong
[2]Machine Learning Department
Carnegie Mellon University, Pittsburgh, PA 15213
{bshi,wlam,yqxu}@se.cuhk.edu.hk
lbing@cs.cmu.edu

Abstract

News reader comments found in many on-line news websites are typically massive in amount. We investigate the task of **C**ultural-common **T**opic **D**etection (CTD), which is aimed at discovering common discussion topics from news reader comments written in different languages. We propose a new probabilistic graphical model called MCTA which can cope with the language gap and capture the common semantics in different languages. We also develop a partially collapsed Gibbs sampler which effectively incorporates the term translation relationship into the detection of cultural-common topics for model parameter learning. Experimental results show improvements over the state-of-the-art model.

1 Introduction

Nowadays the rapid development of information and communication technology enables more and more people around the world to engage in the movement of globalization. One effect of globalization is to facilitate greater connections between people bringing cultures closer than before. This also contributes to the convergence of some elements of different cultures (Melluish, 2014). For example, there is a growing tendency of people watching the same movie, listening to the same music, and reading the news about the same event. This kind of cultural homogenization brings the emergence of commonality of some aspects of different cultures worldwide. It would be beneficial to identify such common aspects among cultures. For example, it can provide some insights for global market and international business (Cavusgil et al., 2014).

Many news websites from different regions in the world report significant events which are of interests to people from different continents. These websites also allow readers around the world to give their comments in their own languages. The volume of comments is often enormous especially for popular events. In a news website, readers from a particular culture background tend to write comments in their own preferred languages. For some important or global events, we observe that readers from different cultures, via different languages, express common discussion topics. For instance, on March 8 2014, Malaysia Airlines Flight MH370, carrying 227 passengers and 12 crew members, disappeared. Upon the happening of this event, many news articles around the world reported it and many readers from different continents commented on this event. Through analyzing the reader comments manually, we observe that both English-speaking and Chinese-speaking readers expressed in their corresponding languages their desire for praying for the MH370 flight. This is an example of a **cultural-common** discussion topic. Identifying such cultural-common topics automatically can facilitate better understanding and organization of the common concerns or interests of readers with different language background. Such technology can be deployed for developing various applications. One application is to build a reader comment digest system that can organize comments by cultural-common discussion topics and rank the topics by popularity. This provides a functionality of analyzing the common focus of readers from different cultures on a particular event. An example of such application is shown in Figure 3. Under each event, reader comments are grouped by cultural-common topics.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 676–685,
Berlin, Germany, August 7-12, 2016.

In this paper, we investigate the task of **C**ultural-common **T**opic **D**etection (CTD) on multilingual news reader comments. Reader comments about a global event, written in different languages, from different news websites around the world exist in massive amount. The main goal of this task is to discover cultural-common discussion topics from raw multilingual news reader comments for a news event. One challenge is that the discussion topics are unknown. Another challenge is related to the language gap issue. Precisely, the words of reader comments in different languages are composed of different terms in their corresponding languages. Such language gap issue poses a great deal of challenge for identifying cultural-common discussion topics in multilingual news comments settings.

One recent work done by Prasojo et al. (2015) is to organize news reader comments around entities and aspects discussed by readers. Such organization of reader comments cannot handle the identification of common discussion topics. On the other hand, the Muto model proposed by Boyd-Graber and Blei (2009) can extract common topics from multilingual documents. This model merely outputs cross-lingual topics of matching word pairs. One example of such kind of topic contains key terms of word pairs such as "plane:飞机 ocean:海洋 ...". The assumption of one-to-one mapping of words has some drawbacks. One drawback is that the correspondence of identified common topics is restricted to the vocabulary level. Another drawback is that the one-to-one mapping of words cannot fit the original word occurrences well. For example, the English term "plane" appears in the English documents frequently while the Chinese translation "飞机" appears less. It is not reasonable that "plane" and "飞机" share the same probability mass in common topics. Another closely related existing work is the PCLSA model proposed by Zhang et al. (2010). PCLSA employs a mixture of English words and Chinese words to represent common topics. It incorporates bilingual constraints into the Probabilistic Latent Semantic Analysis (PLSA) model (Hofmann, 2001) and assumes that word pairs in the dictionary share similar probability in a common topic. However, similar to one-to-one mapping of words, such bilingual constraints cannot handle well the original word co-occurrence in each language resulting in a degradation of the coherence and interpretability of common topics.

We propose a new probabilistic graphical model which is able to detect cultural-common topics from multilingual news reader comments in an unsupervised manner. In principle, no labeled data is needed. In this paper, we focus on dealing with two languages, namely, English and Chinese news reader comments. Different from prior works, we design a technique based on auxiliary distributions which incorporates word distributions from the other language and can capture the common semantics on the topic level. We develop a partially collapsed Gibbs sampler which decouples the inference of topic distribution and word distribution. We also incorporate the term translation relationship, derived from a bilingual dictionary, into the detection of cultural-common topics for model parameter learning.

We have prepared a data set by collecting English and Chinese reader comments from different regions reflecting different culture. Our experimental results are encouraging showing improvements over the state-of-the-art model.

2 Related Work

Prasojo et al. (2015) and Biyani et al. (2015) organized news reader comments via identified entities or aspects. Such kind of organization via entities or aspects cannot capture common topics discussed by readers. Digesting merely based on entities fails to work in multilingual settings due to the fact that the common entities have distinct mentions in different languages.

Zhai et al. (2004) discovered common topics from comparable texts via a PLSA based mixture model. Paul and Girju (2009) proposed a Mixed-Collection Topic Model for finding common topics from different collections. Despite the fact that the above models can find a kind of common topic, they only deal with a single language setting without considering the language gap.

Some works discover common latent topics from multilingual corpora. For aligned corpora, they assume that the topic distribution in each document is the same (Vulić et al., 2011; Vulić and Moens, 2014; Erosheva et al., 2004; Fukumasu et al., 2012; Mimno et al., 2009; Ni et al., 2009; Zhang et al., 2013; Peng et al., 2014). However, aligned corpora are often unavailable for most domains. For unaligned corpora, cross-lingual topic models use some language resources, such

as a bilingual dictionary or a bilingual knowledge base to bridge the language gap (Boyd-Graber and Blei, 2009; Zhang et al., 2010; Jagarlamudi and Daumé III, 2010). As mentioned above, the goals of Boyd-Graber and Blei (2009) as well as Jagarlamud and Daumé (2010) focus on mining the correspondence of topics at the vocabulary level, which are different from that of Zhang et al. (2010) and ours. The model in Zhang et al. (2010) adds the constraints of word translation pairs into PLSA. These constraints cannot handle the original word co-occurrences well. In contrast, we consider the language gap by incorporating word distributions from the other language, capturing the common semantics on the topic level. Moreover, we use a fully Bayesian paradigm with a prior distribution.

Some existing topic methods conduct cross-lingual sentiment analysis (Lu et al., 2011; Guo et al., 2010; Lin et al., 2014; Boyd-Graber and Resnik, 2010). These models are not suitable for our CTD task because they mainly detect common elements related to product aspects. Moreover some works focus more on detecting sentiments.

3 Our Proposed Model

3.1 Model Description

The problem definition of the CTD task is described as follows. For a particular event, both English and Chinese news reader comments are collected from different regions reflecting different culture. The set of English comments is denoted by $\mathcal{E}$ and the set of Chinese comments is denoted by $\mathcal{C}$. The goal of the CTD task is to extract cultural-common topics $k \in \{1, 2, \ldots, K\}$ from $\mathcal{E}$ and $\mathcal{C}$. The set of multilingual news reader comments of each event are processed within the same event.

Our proposed model is called **M**ultilingual **C**ultural-common **T**opic **A**nalysis (MCTA) which is based on graphical model paradigm as depicted in Figure 1. The plate on the right represents cultural-common topics. Each cultural-common topic k is represented by an English word distribution φ_k^e over English vocabulary Λ^e and a Chinese word distribution φ_k^c over Chinese vocabulary Λ^c. We make use of a bilingual dictionary, which is composed of many-to-many word translations among English and Chinese words. To capture common semantics of multilingual news reader comments, we design two auxiliary distributions η_e, with dimension Λ^e, and η_c, with dimension Λ^c, to help the generation of φ_k^e and φ_k^c. Precisely, we generate η_e and η_c from the Dirichlet prior distributions $Dir(\beta \cdot \mathbf{1}_{|\Lambda^e|})$ and $Dir(\beta \cdot \mathbf{1}_{|\Lambda^c|})$ respectively, where $\mathbf{1}_D$ denotes a D-dimensional vector whose components are 1. Then we draw φ_k^e from the mixture of η_k^e and the translation of η_k^c. It is formulated as:

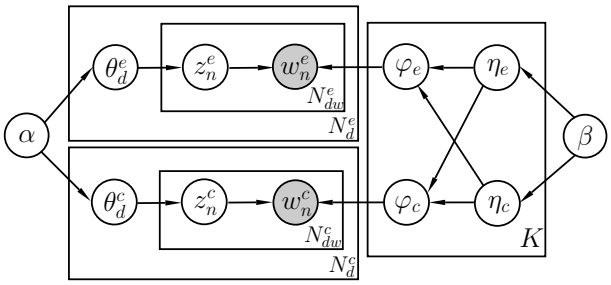

Figure 1: Our proposed graphical model

$$\varphi_k^e \propto \lambda(\eta_k^c)^T \mathbf{M}^{c \to e} + (1-\lambda)\eta_k^e \quad \text{where} \quad \eta^e, \eta^c \sim Dir(\beta) \tag{1}$$

where $\lambda \in (0, 1)$ is a parameter which balances the nature of original topics and transferred information from the other language. $\mathbf{M}^{c \to e}$ is a mapping $|\Lambda^c| \times |\Lambda^e|$ matrix from Λ^c to Λ^e. Each element $M_{ij}^{c \to e}$ is the mapping occurrence probability of the English term w_j^e given the Chinese term w_i^c in the set of news reader comments. This probability is calculated as:

$$M_{ij}^{c \to e} = \frac{C(w_j^e) + 1}{|T(w_i^c)| + \sum_{w^e \in T(w_i^c)} C(w^e)} \tag{2}$$

where $C(w_j^e)$ is the count of w_j^e in all news reader comments and $T(w_i^c)$ is the set of English translations of w_i^c found in the bilingual dictionary. The "add-one" smoothing is adopted. Note that the sum of each row is equal to 1. Using the same principle, we can derive φ_k^c which can be formulated as:

$$\varphi_k^c \propto \lambda(\eta_k^e)^T M^{e \to c} + (1-\lambda)\eta_k^c \quad \text{where} \quad \eta^e, \eta^c \sim Dir(\beta) \tag{3}$$

As a result, the incorporation of η_k^e and η_k^c on the topic level encourages the word distribution φ_k^e and φ_k^c to share common semantic components of reader comments in different languages.

The upper left plate in Figure 1 represents English reader comments. N_d^e denotes the number of English reader comments and N_{dw}^e denotes

the number of words in the English comment d^e. Each English reader comment d^e is characterized by a K-dimensional topic membership vector θ_d^e, which is assumed to be generated by the prior $Dir(\alpha \cdot \mathbf{1}_K)$. For each word w_n^e in an English comment d^e, we generate the topic z_e^n from θ_d^e. We generate the word w_n^e from the corresponding distribution φ_k^e.

The bottom left plate in Figure 1 represents Chinese reader comments. Similarly, we generate the topic distribution θ_d^c from the prior $Dir(\alpha \cdot \mathbf{1}_K)$. The topic z_c^n of each word w_n^c in a Chinese comment d^c is generated from θ_d^c. We generate word w_n^c from the corresponding distribution φ_k^c.

The generative process is formally depicted as:

- For each topic $k \in \mathcal{K}$
  - choose auxiliary distributions $\eta_k^e \sim Dir(\beta \cdot \mathbf{1}_{|\Lambda^e|})$ and $\eta_k^c \sim Dir(\beta \cdot \mathbf{1}_{|\Lambda^c|})$
  - choose English word distribution φ_k^e and φ_k^c using Eq. 1 and Eq. 3 respectively.
- For each English comment $d^e \in \mathcal{E}$, choose $\theta_d^e \sim Dir(\alpha \cdot \mathbf{1}_K)$
  - For each position n in d^e
    - draw $z_n^e \sim$ Multi(θ_d^e)
    - draw $w_n^e \sim$ Multi(φ_k^e)
- For each Chinese comment $d^c \in \mathcal{C}$, choose $\theta_d^c \sim Dir(\alpha \cdot \mathbf{1}_K)$
  - For each position n in d^c
    - draw $z_n^c \sim$ Multi(θ_d^c)
    - draw $w_n^c \sim$ Multi(φ_k^c)

Note that for simplicity, we present our model on the bilingual setting of Chinese and English. It can be extended to multilingual setting via introducing auxiliary distributions for each language. Each topic word distribution for each language is generated by the convex combination of all the auxiliary distributions.

3.2 Posterior Inference

In order to decouple the inference of z_n and φ_k for each language, we develop a partially collapsed Gibbs method which just discards θ_d^e and θ_d^c. Given φ_k^e, we sample the new assignments of the topic z_{di}^e in English news reader comments d^e with the following conditional probability:

$$P(z_{di}^e = k|\mathbf{z}^{e,\neg i}, W^e, \alpha, \varphi_k^e) \propto (N_{dk}^{e,\neg i} + \alpha_k) \times \varphi_k^e \tag{4}$$

where $\mathbf{z}^{e,\neg i}$ denotes the topic assignments except the assignment of the ith word. N_{dk}^e is the number of words in English document d^e whose topics are assigned to k. Similarly, we sample z_{di}^c with the following equation:

$$P(z_{di}^c = k|\mathbf{z}^{c,\neg i}, W^c, \alpha, \varphi_k^c) \propto (N_{dk}^{c,\neg i} + \alpha_k) \times \varphi_k^c \tag{5}$$

Given the topic assignments, the probability of the entire comment set can be:

$$p(W|\mathbf{z}, \varphi_k^e, \varphi_k^c) = \prod_{w \in \Lambda^e} (\varphi_{kw}^e)^{N_{kw}^e} \times \prod_{w \in \Lambda^c} (\varphi_{kw}^c)^{N_{kw}^c} \tag{6}$$

where N_{kw}^e is the number of words w in English news reader comments assigned to the topic k and N_{kw}^c is the number of words w in Chinese news reader comments assigned to the topic k.

Using Eq. 6, we can obtain the posterior likelihood related to η_k^e and η_k^c:

$$\begin{aligned}\mathcal{L}_{MAP} = & \\ & \sum_{w_i \in \Lambda^e} N_{kw_i}^e \log(\lambda \sum_{w_j \in \Lambda^c} M_{ji}^{c \to e} \eta_{kw_j}^c + (1-\lambda)\eta_{kw_i}^e) \\ & + \sum_{w_i \in \Lambda^c} N_{kw_i}^c \log(\lambda \sum_{w_j \in \Lambda^e} M_{ji}^{e \to c} \eta_{kw_j}^e + (1-\lambda)\eta_{kw_i}^c) \\ & + \sum_{w_i \in \Lambda^e} (\beta - 1) \log \eta_{kw_i}^e + \sum_{w_i \in \Lambda^c} (\beta - 1) \log \eta_{kw_i}^c\end{aligned} \tag{7}$$

We optimize Eq. 7 under the constraints of $\sum_{w_i \in \Lambda^e} \eta_{kw_i}^e = 1$ and $\sum_{w_i \in \Lambda^c} \eta_{kw_i}^c = 1$. Using the fixed-point method, we obtain the update

Algorithm 1 Partially Collapsed Gibbs Sampling for MCTA

```
Initialize z, φ_k^e, φ_k^c, η_k^e, η_k^c
for iter = 1 to Max_iter do
    for each English comment d in E do
        for each word w_n^e in d do
            draw z_n^e using Eq. 4
        end for
    end for
    for each Chinese comment d in C do
        for each word w_n^c in d do
            draw z_n^c using Eq. 5.
        end for
    end for
    Update η_k^e, η_k^c by Eq. 8 and Eq. 9
    Update φ_k^e, φ_k^c according to Eq. 1 and Eq. 3
end for
Output θ_dk by Eq. 10
```

equations of $\eta^e_{kw_t}$ and $\eta^c_{kw_t}$ shown in Eq. 8 and Eq. 9.

$$\eta^e_{kw_t} \propto [\frac{(1-\lambda)N^e_{kw_t}}{\lambda \sum_{w_j \in \Lambda^c} M^{c\to e}_{jt} \eta^c_{kw_j} + (1-\lambda)\eta^e_{kw_t}} + \sum_{w_i \in \Lambda^c} \frac{\lambda N^c_{kw_i} M^{e\to c}_{ti}}{\lambda \sum_{w_j \in \Lambda^e} M^{e\to c}_{ji} \eta^e_{kw_j} + (1-\lambda)\eta^c_{kw_i}}]\eta^e_{kw_t} + \beta \tag{8}$$

$$\eta^c_{kw_t} \propto [\frac{(1-\lambda)N^c_{kw_t}}{\lambda \sum_{w_j \in \Lambda^e} M^{e\to c}_{jt} \eta^e_{kw_j} + (1-\lambda)\eta^c_{kw_t}} + \sum_{w_i \in \Lambda^e} \frac{\lambda N^e_{kw_i} M^{c\to e}_{ti}}{\lambda \sum_{w_j \in \Lambda^c} M^{c\to e}_{ji} \eta^c_{kw_j} + (1-\lambda)\eta^e_{kw_i}}]\eta^c_{kw_t} + \beta \tag{9}$$

Moreover, the posterior estimates for the topic distribution θ_d can be computed as follows.

$$\theta_{dk} = \frac{N_{dk} + \alpha}{\sum_{k \in \mathcal{K}} N_{dk} + K\alpha} \tag{10}$$

The whole detailed algorithm is depicted in Algorithm 1. When $\lambda = 0$, the updated equations of η^e_k and η^c_k can be simplified as:

$$\begin{aligned} \eta^e_{kw_t} &\propto N^e_{kw_t} + \beta \\ \eta^c_{kw_t} &\propto N^c_{kw_t} + \beta \end{aligned} \tag{11}$$

Then we have:

$$\begin{aligned} \varphi^e_k &\sim Dir(N^e_{kw_1} + \beta, N^e_{kw_2} + \beta, \dots) \\ \varphi^c_k &\sim Dir(N^c_{kw_1} + \beta, N^c_{kw_2} + \beta, \dots) \end{aligned} \tag{12}$$

Therefore, the algorithm degrades to a Gibbs sampler of LDA.

4 Experiments

4.1 Data Set and Preprocessing

We have prepared a data set by collecting English and Chinese comments from different regions reflecting different culture for some significant events as depicted in Table 1. The English reader comments are collected from Yahoo[1] and the Chinese reader comments are collected from Sina News[2]. We first remove news reader comments whose length is less than 5 words. We remove the punctuations and the stop words. For English comments, we also stem each word to its root form using Porter Stemmer (Porter, 1980). For the Chinese reader comments, we use the Jieba package[3] to segment and remove Chinese stop words. We utilize an English-Chinese dictionary from MDBG[4].

Event	Title	#English comments	#Chinese comments
1	MH370 flight accident	8608	5223
2	ISIS in Iraq	6341	3263
3	Ebola occurs	2974	1622
4	Taiwan Crashed Plane	6780	2648
5	iphone6 publish	5837	4352
6	Shooting of Michael Brown	17547	3693
7	Charlie Hebdo shooting	1845	551
8	Shanghai stampede	3824	3175
9	Lee Kuan Yew death	2418	1534
10	AIIB foundation	7221	3198

Table 1: The statistics for the data set

4.2 Comparative Methods

The PCLSA model proposed by Zhang et al. (2010) can be regarded as the state-of-the-art model for detecting latent common topics from multilingual text documents. We implemented PCLSA as one of the comparative methods in our experiments.

Another comparative model used in the experiment is LDA (Blei et al., 2003), which can generate K English topics and K Chinese topics from English and Chinese reader comments respectively. Then we translate Chinese topics into English topics and use symmetric KL divergence to align translated Chinese topics with original English topics. Each aligned topic pair is regarded as a cultural-common topic.

4.3 Experiment Settings

For each event, we partitioned the comments into a subset of 90% for the graphical model parameter estimation. The remaining 10% is used as a holdout data for the evaluation of the *CCP* metric as discussed in Section 4.4.1. We repeated the runs five times. For each run, we randomly split the comments to obtain the holdout data. As a result, we have five runs for our method as well as comparative methods. We make use of the holdout data of one event, namely the event "MH370

[1] http://news.yahoo.com

[2] http://news.sina.com.cn/world/

[3] https://github.com/fxsjy/jieba

[4] http://www.mdbg.net/chindict/chindict.php?page=cc-cedict

Flight Accident", to estimate the number of topics K for all models and λ in Eq. 1 for our model. The setting of K is described in Section 4.4.3. We set $\lambda = 0.5$ after tuning. For hyper-parameters, we set α to 0.5 and β to 0.01. When performing our Gibbs algorithm, we set the maximum iteration number as 1000, and the burn-in sweeps as 100.

4.4 Cultural-common Topic Evaluation

We conduct quantitative experiments to evaluate how well our MCTA model can discover cultural-common topics.

4.4.1 Evaluation Metrics

We use two metrics to evaluate the topic quality. The first metric is the "cross-collection perplexity" measure denoted as *CCP* which is similar to the one used in Zhang et al. (2010). The *CCP* of high quality cultural-common topics should be lower than those topics which are not shared by the English and Chinese reader comments. The calculation of *CCP* consists of two steps: *1)* For each $k \in \mathcal{K}$, we translate φ_k^e into Chinese word distribution $T(\varphi_k^e)$ and translate φ_k^c English word distribution $T(\varphi_k^c)$. To translate φ_k^e and φ_k^c, we look up the bilingual dictionary and conduct word-to-word translation. If one word has several translations, we distribute its probability mass equally to each English translation. *2)* We use $T(\varphi_k^e)$ to fit the holdout Chinese comments $\mathcal{C}$ and $T(\varphi_k^c)$ to fit the holdout English comments $\mathcal{E}$ using Eq. 13 (Blei et al., 2003). Eq. 13 depicts the calculation of *CCP*. The lower the *CCP* value is, the better the performance is.

$$
\begin{aligned}
&CCP = \\
&\frac{1}{2}\exp\{-\frac{\sum_{d\in\mathcal{E}}\sum_{w\in d}\sum_{k\in\mathcal{K}}\log p(k|\theta_d)p(w|T(\varphi_k^c))}{\sum_{d\in\mathcal{E}} N_d^e}\}+ \\
&\frac{1}{2}\exp\{-\frac{\sum_{d\in\mathcal{C}}\sum_{w\in d}\sum_{k\in\mathcal{K}}\log p(k|\theta_d)p(w|T(\varphi_k^e))}{\sum_{d\in\mathcal{C}} N_d^c}\}
\end{aligned}
\tag{13}
$$

For each detected common topic, we wish to evaluate the degree of commonality. We design another metric called "topic commonality distance" denoted by *TCD*. We first evaluate the KL-divergence between the English topic and translated Chinese topic. We also evaluate the KL-divergence between the Chinese topic and translated English topic. Then *TCD* is computed as the average sum of the two KL-divergences. The lower the *TCD* measure is, the better the topic is.

Event	LDA	PCLSA	MCTA
1	1963.57	1842.24	1784.05
2	1940.03	1831.55	1756.92
3	1958.09	1905.43	1808.01
4	1916.49	1847.16	1775.32
5	1901.44	1797.92	1744.07
6	1916.70	1853.66	1786.77
7	1945.22	1897.15	1824.10
8	1942.29	1862.14	1749.43
9	1943.53	1856.70	1739.66
10	1866.23	1815.44	1749.49
avg.	1929.36	1850.94	**1771.78**

Table 2: Topic quality evaluation as measured by *CCP*

The topic detected by PCLSA is a mixture of English and Chinese words. We obtain English representation and Chinese representation of the topic by the conditional probabilities as given in Eq. 14.

$$
\begin{aligned}
p(w^e|\varphi_k^e) &= \frac{p(w^e|\varphi_k)}{\sum_{w\in\Lambda^e} p(w|\varphi_k)} \\
p(w^c|\varphi_k^c) &= \frac{p(w^c|\varphi_k)}{\sum_{w\in\Lambda^c} p(w|\varphi_k)}
\end{aligned}
\tag{14}
$$

4.4.2 Experimental Results

The average *CCP* values of the three models are shown in Table 2. Our MCTA model achieves the best performance compared with PCLSA and LDA. Both MCTA and PCLSA achieve a better *CCP* than LDA because they can bridge the language gap in the multilingual news reader comments to some extent. Compared with PCLSA, our MCTA model demonstrates a 4.2% improvement. Our MCTA model provides a better characterization of the collections. One reason is that our MCTA model learns the word distribution of cultural-common topics using an effective topic modeling with a prior Dirichlet distribution. It is similar to the advantage of LDA over PLSA. Moreover, the bilingual constraints in PCLSA cannot handle the original natural word co-occurrence well in each language. In contrast, MCTA represents cultural-common topics as a mixture of the original topics and the translated topics, which capture the comment semantics more effectively.

The average *TCD* of three models are shown in Table 3. Our MCTA outperforms the two comparative methods. The cultural-common topics iden-

Event	LDA	PCLSA	MCTA
1	0.029	0.0075	0.0042
2	0.029	0.0072	0.0043
3	0.033	0.0076	0.0046
4	0.031	0.0075	0.0046
5	0.033	0.0086	0.0069
6	0.029	0.0066	0.0058
7	0.036	0.0080	0.0044
8	0.033	0.0079	0.0034
9	0.034	0.0088	0.0036
10	0.029	0.0067	0.0036
avg.	0.032	0.0076	**0.0045**

Table 3: Topic quality evaluation as measured by *TCD*

Event	LDA	PCLSA	MCTA
1	0.128	0.117	0.138
2	0.144	0.126	0.158
3	0.122	0.117	0.120
4	0.138	0.138	0.169
5	0.128	0.109	0.152
6	0.134	0.138	0.152
7	0.103	0.108	0.111
8	0.110	0.099	0.124
9	0.080	0.085	0.096
10	0.138	0.133	0.154
avg.	0.122	0.117	**0.137**

Table 4: Topic coherence evaluation

tified by MCTA have better topic commonality because our MCTA model can capture the common semantics between news reader comments in different languages.

4.4.3 Determining Number of Topics

As mentioned in Section 4.3, we use the hold-out data of one event to determine K. For each $\lambda \in \{0.2, 0.5, 0.8\}$, we vary K in the range of $[5, 200]$. Figure 2 depicts the effect of K on the cross-collection perplexity as measured by *CCP*. We can see that *CCP* decreases with the increase of the number of topics. Moreover, through manual inspection we observed that when K is 30 or more, even though *CCP* decreases, the topics will be repeated. Similar observations for the number of topics can be found in Paul and Girju (2009). Therefore, we set $K = 30$. We can also see that our model is not very sensitive to the balance parameter λ.

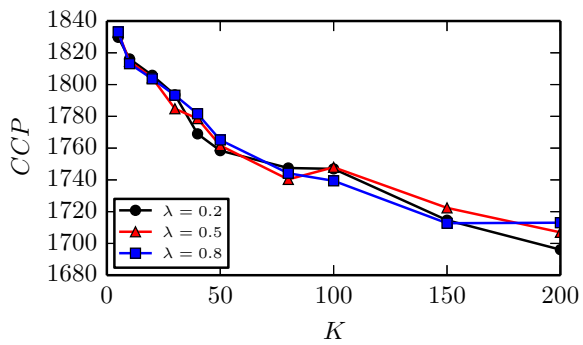

Figure 2: The effect of K

4.5 Topic Coherence Evaluation

We also evaluate the coherence of topics generated by PCLSA and MCTA, which indicates the interpretability of topics. Following Newman et al. (2010), we use a pointwise mutual information (PMI) score to measure the topic coherence. We compute the average PMI score of top 20 topic word pairs using Eq. 15. Newman et al. (2010) observed that it is important to use an external data set to evaluate PMI. Therefore, we use a 20-word sliding window in Wikipedia (Shaoul, 2010) to identify the co-occurrence of word pairs.

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \qquad (15)$$

The experimental results are shown in Table 4. We can see that our MCTA model generally improves the coherence of the learned topics compared with PCLSA. The word-to-word bilingual constraints in PCLSA are not as effective. On the other hand, our MTCA model incorporates the bilingual translations using auxiliary distributions which incorporate word distributions from the other language on the topic level and can capture common semantics of multilingual reader comments.

5 Application and Case Study

We present an application for news comment digest and show some examples of detected cultural-common discussion topics in Figure 3. Under each event, the system can group reader comments into cultural-common discussion topics which can capture common concerns of readers in different languages. For each common topic, it shows top ranked words and corresponding reader comments

Event: MH370 Flight Accident	
Topic Terms	Reader Comments
family hope love dead people victim passenger sad sorry life 家庭(family) 家属(family) 家人(family) 亲人(family) 失望(disappoint) 希望(hope) 心酸(sad) 愿(wish) 痛(pain) 心痛(sad)	I feel sorry for the families of the victims of this flight - this aircraft piece being found probably brings that terrible day back
	I feel so sorry for the relatives of the missing passengers who are doomed to spend the rest of their lives getting their hopes continuously...
	The family members should now begin to have a closure as the plane's flaperon has been identified. The Australians have been proven correct as...
	时间真快，一年半的时间过去了，不知那些失去亲人的朋友们走出悲痛了没有？唯愿逝者在天堂安息，生者在人间安康！
	家属朋友们，失去亲人的痛苦的，但生活是美好的，一定要好好生活，让逝者安心！
	每一次的提起370 就会让那些失去亲人的家人心痛折磨一次！
	⋮
ocean island search India mile area locate Australia drift west 洋流(ocean currents) 印度洋(Indian ocean) 区域(Region) 海里(mile) 搜索(search) 搜寻(search) 搜救(rescure) 海底(sea floor) 澳洲(Australia) 海域(sea area)	They were looking in the West Australian current. That would have brought the part to the north of Australia. If it got into I equatorial current...
	They need to start their sonar scans about 1000 miles south of the tip of India seeing how the currents in that ocean work, and how long it took for that piece to float to the island so far out. It's pretty simple to estimate seeing how Fukushima fishing boats travelled a set distance over a set time, given a set current...
	look at current maps. well off the western coast of Aus is the S. Equitorial Current in the Indian Ocean which flows in a circular counter clockwise pattern. It most certainly could have come from a plane that crashed off the AUS coast.
	这么多阴谋论者说这是美国搞的鬼，我只能呵呵了，美国的调查结论说是在南印度洋，在澳大利亚那边，现在发现的位置是不是和美国的调查结果一致？洋流的运动方向和推测地点、残骸地点是否符合？为什么一定要把空难说成某国的阴谋才甘心？
	南印度洋的洋流是自东向西，这个残骸落在这里，那么飞机应该坠毁在东方的海面上。
	即使这片残骸属于MH370客机，在留尼汪被发现也并不意味着飞机的失事地点就在留尼汪。假设飞机在澳大利亚海域坠毁，其残骸很有可能被洋流带到印度洋，一年以后被海浪冲上安德烈海滩.
	⋮
⋮	

Event:ISIS in Iraq	
Topic Terms	Reader Comments
muslim islam religion world christian god people believe jew human 信仰(belief) 宗教(religion) 世界(world) 相信(believe) 全世界(world) 伊斯兰(Islam) 穆斯林(muslim) 人(people) 犹太(jew) 人类(human)	1 don't understand Muslims, Islam or the Holy Qur'an! The aim of Islam is not to instil Sharia over the entire world, Islam preaches that you believe in God worship Him alone and do right good by your belief.....
	Oh, I get it. It's about the badness of Muslims being humbled and humiliated in prison by Americans. But IS rapes and mutilates and pilliages...
	If there was no Muslim religion in Iraq, there would be no ISIS because there would have been no necessity for a thug like Saddam to control...
	ISIS是个宗教极端组织没错，但是如果ISIS没有下这个令，而是被故意栽赃，其用意显然不在针对ISIS本身
	宗教不是祸首，真正的魁首是打着宗教旗号的极端分子，都说国人没信仰但是一样有右翼激进分子，激进的民族主义，然后是民粹，最后是种族主义，纳粹不也是这么一步一步上台的么，历史总是似曾相识。
	可以看出一切邪恶都是出自宗教！宗教欺骗人类的另一面！以后别拿我们汉族没有信仰来说事了！看看你们信仰的后果吧！
	⋮
⋮	

Event:AIIB Foundation	
Topic Terms	Reader Comments
bank aiib world imf asian develop investment institution infrastructure member 银行(bank) 金融(finance) 世界银行(The World Bank) 世行(The World Bank) 金融机构(Finance institution) 亚洲开发银行(Asian Development Bank) 成员(member) 国际货币基金组织(IMF) 国际(International) 世界(world)	Looks like all the rats are jumping off the sinking world bank and IMF ship. America has pushed their bulling ways long enough and people are...
	The Federal Reserve, the World Bank, The IMF, and the BIS are failed, self-serving institutions. One can only hope that China will stimulate world growth and the suppression by the west will finally come to an end. The US dollar no longer deserves to be the world's reserve currency.
	Bank shopping !!! No more stranglehold by the IMF and World Bank. If Ukraine had only waited another year. Too bad.
	把米国和日本排除在亚投行之外，让他们自己单独经营亚洲开发银行和世界银行！[哈哈]
	欧洲国家被美国坑惨了，世界银行、国际货币基金的钱都在为美国服务，反过来美国又利用乌克兰危机打压欧元，如今欧洲国家看明白了，还是中国靠普。
	世界银行和亚洲开发银行都对亚投行表示欢迎，明显有些言不由衷。信他才怪
	⋮

Figure 3: Some sample common discussion topics of some events

according to θ^e_{dk} and θ^c_{dk}. Considering the event "MH370 Flight accident", it shows two of the detected cultural-common topics. The first one indicates that readers pray for the family in the accident and the second one is related to the search of the crashed plane. For the common topic about praying for the family, we can see that the topics contain both English words and Chinese words which are very relevant and share common semantics of "family" and "hope". Moreover, the corresponding English and Chinese reader comments, both of which mention the family in the accident, illustrate a high coherent common discussion topic. Similarly for the second common topic, there are common semantics between English and Chinese top ranked words about the search of the crashed plane. Some of the English comments and Chinese comments mention the query of the position of the crashed plane. Interesting common topics are also generated for other events, such as the common topic of religion for the event "ISIS in Iraq" and the topic of economic organization for the event "AIIB foundation".

6 Conclusions

We investigate the task of cultural-common discussion topic detection from multilingual news reader comments. To tackle the task, we develop a new model called MCTA which can cope with the language gap and extract coherent cultural-common topics from multilingual news reader comments. We also develop a partially collapsed Gibbs sampler which incorporates the term translation relationship into the detection of cultural-common topics effectively for model parameter learning.

Acknowledgments

The work described in this paper is substantially supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14203414) and the Direct Grant of the Faculty of Engineering, CUHK (Project Code: 4055034). This work is also affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. We also thank the anonymous reviewers for their insightful comments.

References

Prakhar Biyani, Cornelia Caragea, and Narayan Bhamidipati. 2015. Entity-specific sentiment classification of yahoo news comments. *arXiv preprint arXiv:1506.03775*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David M Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 75–82.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 45–55.

S Tamer Cavusgil, Gary Knight, John R Riesenberger, Hussain G Rammal, and Elizabeth L Rose. 2014. *International business: strategy, management and the new realities*. Pearson Australia.

Elena Erosheva, Stephen Fienberg, and John Lafferty. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5220–5227.

Kosuke Fukumasu, Koji Eguchi, and Eric P Xing. 2012. Symmetric correspondence topic models for multilingual text analysis. In *Advances in Neural Information Processing Systems*, pages 1286–1294.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, and Zhong Su. 2010. Opinionit: a text mining system for cross-lingual opinion analysis. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1199–1208.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd European Conference on IR Research*, pages 444–456. Springer.

Zheng Lin, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. 2014. A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 1089–1098.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 320–330.

Steve Melluish. 2014. Globalization, culture and psychology. *International Review of Psychiatry*, 26(5):538–543.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pages 1155–1156.

Michael Paul and Roxana Girju. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1417.

Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 674–679.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Radityo Eko Prasojo, Mouna Kacimi, and Werner Nutt. 2015. Entity and aspect extraction for organizing news comments. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 233–242.

Cyrus Shaoul. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta*.

Ivan Vulić and Marie-Francine Moens. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 349–362.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 479–484.

ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 743–748.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137.

Tao Zhang, Kang Liu, and Jun Zhao. 2013. Cross lingual entity linking with bilingual topic model. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2218–2224.