Bursty Topics in Time Series Japanese / Chinese News Streams and their Cross-Lingual Alignment

Liyi Zheng Takehito Utsuro
Grad. Sch. Sys. & Inf. Eng.,
University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Masaharu Yoshioka
Grad. Sch. Inf. Sci. & Tech.,
Hokkaido University,
Sapporo, 060-0808, JAPAN

Abstract

This paper studies issues regarding topic modeling of information flow in multilingual news streams. If someone wants to find differences in the topics of Japanese news and Chinese news, it is usually necessary for him/her to carefully watch every article in Japanese and Chinese news streams at every moment. In such a situation, topic models such as LDA (Latent Dirichlet Allocation) and DTM (dynamic topic model) are quite effective in estimating distribution of topics over a document collection such as articles in a news stream. To the results of estimating distribution of topics in Japanese / Chinese news streams, we apply Kleinberg's modeling of bursts, and detect bursty topics for both Japanese and Chinese news. Finally, we propose how to cross-lingually align those bursty topics in time series Japanese / Chinese news streams. We also show that, by detecting bursty topics in advance of cross-lingual topic alignment and aligning topics that are bursty for both Japanese and Chinese, correct rate of cross-lingual topic alignment improves.

1 Introduction

Among various types of recent information explosion, that in news stream and blogs is also a kind of serious problems. Especially, recent studies [14, 7, 6, 1, 11] focusing on multilingual information sources such as news and blogs argue various useful perspectives regarding multilingual information sources. For example, Yangarber et al. [14] studied how to combine reports on epidemic threats from over 1,000 portals in 32 languages. Bautin et al. [1] studied how to analyze sentiment distribution in news articles across 9 languages. Fukuhara et al. [7] and Nakasaki et al. [11] studied how to cross-lingually analyze multilingual blogs collected with a topic keyword. In those previous works, how to efficiently discover differences of concerns and opinions on a certain issue was examined. Considering such a motivation, Evans et al. [6] concentrated on developing and evaluating multilingual sentiment analysis techniques, where it has been argued that it is very informative to discover differences of sentiments across languages over a certain concern.

Based on the observation above, this paper studies issues regarding topic modeling of information flow in multilingual news streams. If someone wants to find differences in the topics of Japanese news and Chinese news, it is usually necessary for him/her to carefully watch every article in Japanese and Chinese news streams at every moment. In such a situation, topic models such as LDA (Latent Dirichlet Allocation) [3] and DTM (dynamic topic model) [2] are quite effective in estimating distribution of topics over a document collection such as articles in a news stream. Especially, as a topic model, this paper employs DTM, but not LDA, since it can consider correspondence between topics of consecutive dates. In DTM, we suppose that the data is divided by time slice, for example by date. DTM models the documents (such as articles of news stream) of each slice with a K-component topic model, where the k-th topic at slice t smoothly evolves from the k-th topic at slice $t-1$.

Based on the results of estimating distribution of topics in Japanese / Chinese news streams, this paper proposes how to analyze cross-lingual alignment of topics in time series Japanese / Chinese news streams. Especially, this paper employs Kleinberg's modeling of bursts [9], and applies it to the results of estimating distribution of topics in Japanese / Chinese news streams, in order to detect bursty topics [12, 10] for both Japanese and Chinese news. The overall flow of the proposed framework is illustrated in Figure 1. In order to bridge the gaps between the two languages, namely, Japanese and Chinese, we use Japanese and Chinese term translation pairs extracted from Wikipedia utilizing interlanguage links. With those translation knowledge, we propose how to cross-lingually align those bursty topics in time series Japanese / Chinese news streams. In our previous work [8], we showed that correct rate of cross-lingual topic alignment is about 67$\sim$80% when without bursty topic detection. In this paper, on the other hand, we show that, with the same data set as in our previous work [8], correct rate of cross-lingual topic alignment improves up to 100% by detecting bursty topics in advance of cross-lingual topic alignment and aligning topics that are bursty for both Japanese and Chinese.

Figure 2 shows an example of estimating time series

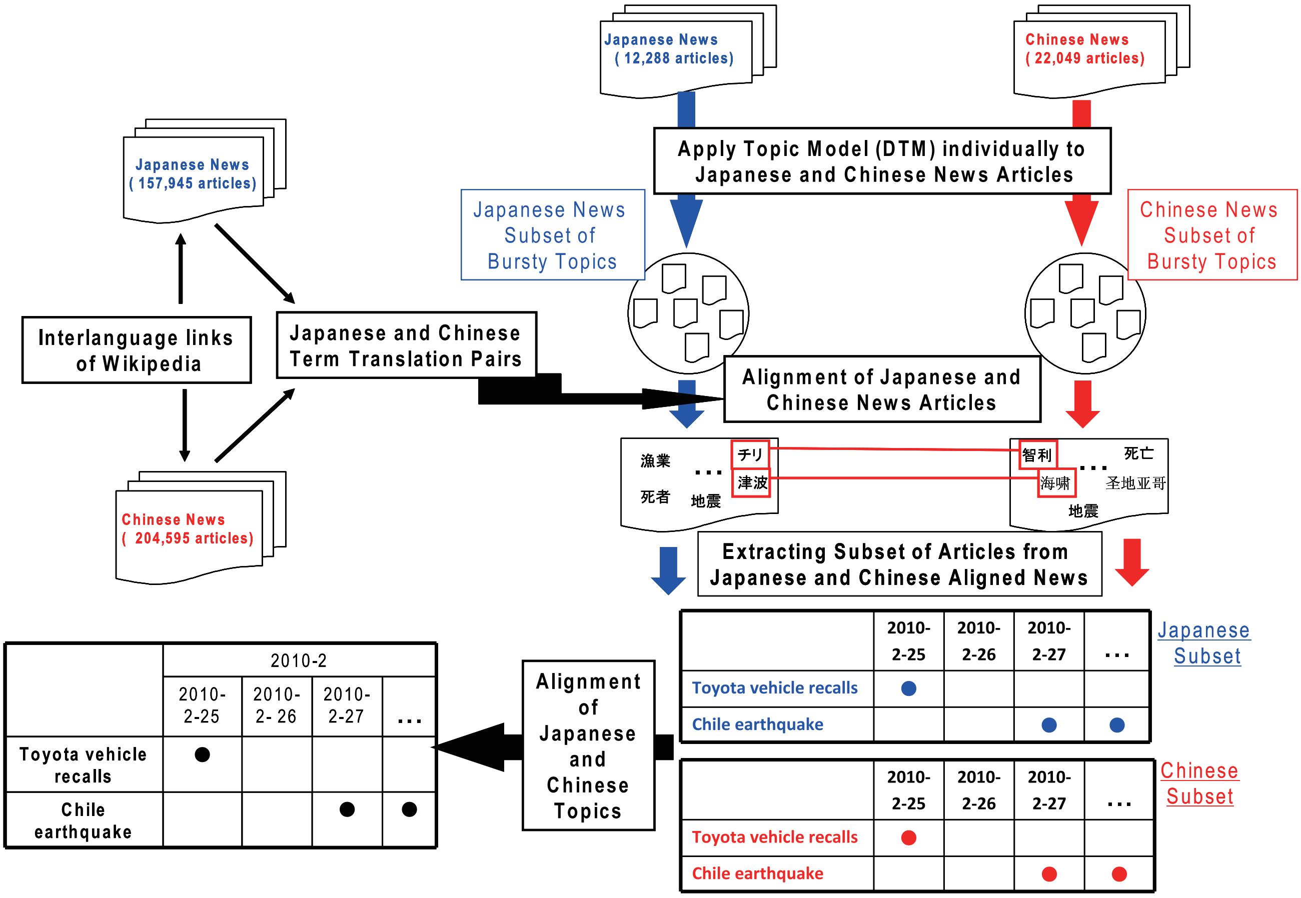
Japanese News
(12,288 articles)
Chinese News
(22,049 articles)
Japanese News
(157,945 articles)
Apply Topic Model (DTM) individually to Japanese and Chinese News Articles
Japanese News Subset of Bursty Topics
Chinese News Subset of Bursty Topics
Interlanguage links of Wikipedia
Japanese and Chinese Term Translation Pairs
Alignment of Japanese and Chinese News Articles
漁業
死者
地震
チリ
津波
智利
海啸
死亡
圣地亚哥
地震
...
Chinese News
(204,595 articles)
Extracting Subset of Articles from Japanese and Chinese Aligned News
2010-2
2010-2-25
2010-2-26
2010-2-27
...
Toyota vehicle recalls
Chile earthquake
●
Alignment of Japanese and Chinese Topics
Japanese Subset
Chinese Subset

Figure 1. Overall Flow of Topic Alignment in Time Series Japanese / Chinese News

topics monolingually for both Japanese and Chinese. The proposed method of cross-lingual topic alignment is successfully applied to those Japanese and Chinese time series news articles, where several topics such as “Toyota vehicle recalls” and “Chile earthquake” are cross-lingually aligned between Japanese and Chinese. Once we have such a cross-lingual topic alignment, it becomes quite easier for us to find certain differences in concerns. For example, in the case of the topic “Chile earthquake”, in Japan, “warn of tsunami” is apparently one of the major concerns, while in Chinese, “emergency assistance was dispatched to Chile” is one of the major concerns.

2 Kleinberg’s Bursts Modeling of a Time Series Topic Model

2.1 Topic Model

As a time series topic model, this paper employs DTM (dynamic topic model) [2]. Unlike LDA (Latent Dirichlet Allocation) [3], in DTM, we suppose that the data is divided by time slice, for example by date. DTM models the documents (such as articles of news stream) of each slice with a K-component topic model, where the k-th topic at slice t smoothly evolves from the k-th topic at slice $t-1$.

In this paper, in order to model time series news stream in terms of a time series topic model, we consider date as the time slice t. Given the number of topics K as well as time series sequence of batches each of which consists of documents represented by a sequence of words w, on each date t (i.e., time slice t), DTM estimated the distribution $p(w \mid z_n)$ $(w \in V)$ of a word w given a topic z_n $(n = 1, \dots, K)$ as well as that $p(z_n \mid d)$ $(n = 1, \dots, K)$ of a topic z_n given a document d, where V is the set of words appearing in the whole document set. In this paper, we estimate the distributions $p(w \mid z_n)$ $(w \in V)$ and $p(z_n \mid d)$ $(n = 1, \dots, K)$ by a Blei’s toolkit[1], where the parameters are tuned through a preliminary evaluation as the number of topics $K = 30$ as well as $\alpha = 0.01$.

2.2 Modeling Bursty Topics in a Topic Model

This section briefly describes the framework [12, 10] of modeling bursty topics among those estimated through the topic modeling framework of the previous section. The framework [12, 10] of modeling bursty topics is based on the modeling of *enumerating bursts* in Kleinberg [9].

[1] http://www.cs.princeton.edu/~blei/topicmodeling.html

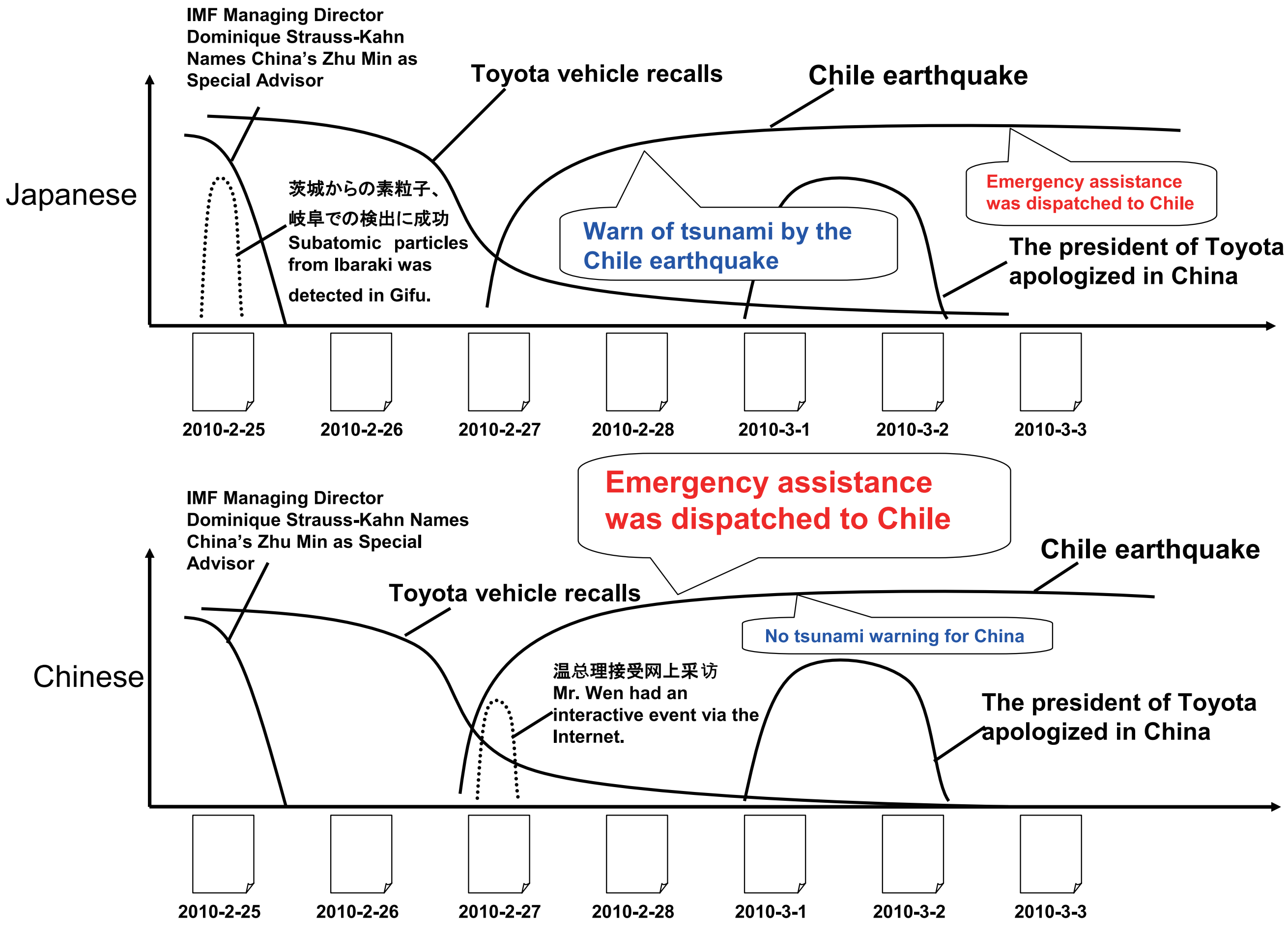

Figure 2. Topic Estimation in Time Series Japanese / Chinese News

Suppose that there are l batches of documents; the t-th batch B_t in the sequence $\mathbf{B} = (B_1, \ldots, B_l)$ of l batches contains r_t relevant documents out of a total of m_t. Here, Kleinberg's modeling of *enumerating bursts* takes as its input the total number m_t of documents in a day (i.e., the t-th batch) and the number r_t of *relevant* documents in a day (i.e., the t-th batch). In Kleinberg's modeling of keyword bursts, a document is simply regarded as *relevant* when containing a particular keyword, and then count the number r_t of relevant documents out of a total of m_t.

In the modeling of topic bursts in Takahashi et al. [12, 10], on the other hand, we first regard a document d as *relevant* to a certain topic z_n that are estimated through the DTM topic modeling procedure, to the degree of the amount of the probability $p(z_n|d)$. We then estimate the number r_t of relevant documents out of a total of m_t simply by summing up the probability $p(z_n|d)$ over the whole document set:

$$r_t = \sum_d p(z_n|d)$$

Once we have the number r_t, then we can estimate the total number of relevant documents throughout the whole batch sequence $\mathbf{B} = (B_1, \ldots, B_l)$ as $R = \sum_{t=1}^{l} r_t$. Having the total number of documents throughout the whole batch sequence as $M = \sum_{t=1}^{l} m_t$, we can estimate the expected fraction of relevant documents as $p_0 = R/M$. Then, by simply following the formalization of keyword bursts presented in Kleinberg [9], it is quite straightforward to model bursty topics in a topic model. As the two parameters s and γ for bursty topic detection[2], we set s as 2.8 and γ as 1 through a preliminary evaluation.

As the results of the DTM topic modeling with Japanese and Chinese news articles in the period of evaluation, K topics are estimated for both Japanese and Chinese for each date. Then, by applying the bursty topic modeling technique, on the i-th day of the period of evaluation, we have the set TT_J^i of bursty Japanese topics, out of the whole K topics. In the similar way, on the i-th day of the period of evaluation, we have the set TT_C^i of bursty Chinese topics, out of the whole K topics.

[2] In the formalization of keyword bursts presented in Kleinberg [9], $p_0 = R/M$ is an expected fraction of relevant documents in the *non-burst* state. On the other hand, the expected fraction p_1 of relevant documents in the *burst* state is introduced as $p_1 = p_0 s$, where we tune the scaling parameter s through a preliminary evaluation. Also, γ is introduced as a cost associated with the state transition from the non-burst state to the burst state, which was again tuned through a preliminary evaluation.

3 Extracting Japanese-Chinese Term Translation utilizing Interlanguage Links in Wikipedia

In this paper, we use Japanese and Chinese term translation pairs extracted from Wikipedia utilizing interlanguage links [8]. More specifically, since we collect Chinese news articles distributed within mainland China which are written in simplified Chinese characters, we extract translation pairs of Japanese terms and simplified Chinese character terms. Its detailed procedure is in Hu et al. [8].

In the evaluation of this paper, we first collect Japanese and Chinese news stream text articles during the period from June 1st, 2009 to May 31st, 2010. In total, 157,945 Japanese news articles are collected from three newspaper companies Yomiuri[3], Nikkei[4], and Asahi[5], while 204,595 Chinese news articles are collected from People's Daily[6]. Then, from the collected news articles, 93,258 Japanese Wikipedia entry titles are collected, out of which 28,071 have interlanguage links to Chinese, while 94,164 Chinese terms in simplified Chinese characters are collected, out of which 28,127 have interlanguage links to Japanese. Finally, from them, 78,519 term translation pairs are collected between Japanese and simplified Chinese characters[7]. The set of those 78,519 term translation pairs is denoted as X_{JC} below:

$$\begin{aligned} X_{JC} &= \Big\{ \langle x_J, x_C \rangle \mid \langle x_J, x_C \rangle \text{ is a term} \\ &\qquad \text{translation pair extracted from} \\ &\qquad \text{Japanese and Chinese news from} \\ &\qquad \text{June 1st, 2009 to May 31st, 2010,} \\ &\qquad \text{and Wikipedia.} \Big\} \end{aligned}$$

Here, we do not explicitly exclude stop words from those term translation pairs, since we assume that the titles of Wikipedia entries with interlanguage links do not tend to be any stop words, but rather specific content words and technical terms.

4 Cross-lingual Topic Alignment

This section proposes the whole framework of cross-lingual topic alignment.

4.1 Cross-Lingual Alignment of News Articles

When cross-lingually aligning Japanese and Chinese news articles, we first count the number of Japanese and Chinese term translation pairs which are shared between the Japanese and Chinese news articles published on the same day. We then align the pair of a Japanese and a Chinese news articles for which the number of shared Japanese and Chinese term translation pairs is more than or equal to the lower bound θ_{JC} (in this paper, θ_{JC} is set as 8 through a preliminary evaluation, where we assume that the length of the news articles we use in the evaluation has a relatively small variance and that it is sufficient to be set as a fixed lower bound regardless of the length of the news articles).

More specifically, on the i-th day, first, for each Japanese topic $t_J (\in$ the set TT_J^i of bursty Japanese topics introduced in section 2.2), we collect news articles d_J which satisfy $P(t_J|d_J) \geq \theta_t$ (in this paper, θ_t is set as 0.6 through a preliminary evaluation). In the similar way, for each Chinese topic $t_C (\in$ the set TT_C^i of bursty Chinese topics introduced in section 2.2), we collect news articles d_C which satisfy $P(t_C|d_C) \geq \theta_t$. Next, given a pair of a Japanese news article d_J and a Chinese news article d_C published on the same day, let $N_{JC}(d_J, d_C)$ be the number of Japanese and Chinese term translation pairs which are shared between d_J and d_C:

$$\begin{aligned} N_{JC}(d_J, d_C) &= \Big| \Big\{ \langle x_J, x_C \rangle (\in X_{JC}) \mid \\ &\qquad x_J \text{ appears in } d_J. \\ &\qquad x_C \text{ appears in } d_C. \Big\} \Big| \end{aligned}$$

Then, for each date, the sets $DD_{JC}(\theta_{JC})$ and $DD_{CJ}(\theta_{JC})$ of pairs of Japanese and Chinese news articles for which the number of shared Japanese and Chinese term translation pairs is more than or equal to the lower bound θ_{JC} are defined as below:

$$\begin{aligned} DD_{JC}(\theta_{JC}) &= \Big\{ \langle d_J, d_C \rangle \mid N_{JC}(d_J, d_C) \geq \theta_{JC}, \\ &\qquad d_C = \operatorname*{argmax}_{d'_C} N_{JC}(d_J, d'_C) \Big\} \\ DD_{CJ}(\theta_{JC}) &= \Big\{ \langle d_J, d_C \rangle \mid N_{JC}(d_J, d_C) \geq \theta_{JC}, \\ &\qquad d_J = \operatorname*{argmax}_{d'_J} N_{JC}(d'_J, d_C) \Big\} \end{aligned}$$

Here, $DD_{JC}(\theta_{JC})$ is created by collecting pairs $\langle d_J, d_C \rangle$, where, for each d_J, d_C is the one with the maximum number N_{JC}. In the similar way, $DD_{CJ}(\theta_{JC})$ is created by collecting pairs $\langle d_J, d_C \rangle$, where, for each d_C, d_J is the one with the maximum number N_{JC}. Note that $DD_{JC}(\theta_{JC})$ is not always the same as $DD_{CJ}(\theta_{JC})$, since the mapping between d_J and d_C is not always one-to-one, but could be one-to-many or many-to-one.

4.2 Cross-Lingual Alignment of Topics

Next, this section proposes how to cross-lingually align topics estimated by a topic model.

First, on the i-th day, given a pair of a Japanese topic $t_J (\in TT_J^i)$ and a Chinese topic $t_C (\in TT_C^i)$ with parameters θ_t and θ_{JC}, we count the number of pairs of collected news articles $\langle d_J, d_C \rangle$ included in $DD_{JC}(\theta_{JC})$ or $DD_{CJ}(\theta_{JC})$, and define $M_{JC}(t_J, t_C, \theta_t, \theta_{JC})$ to be the

[3] http://www.yomiuri.co.jp/

[4] http://www.nikkei.com/

[5] http://www.asahi.com/

[6] http://www.people.com.cn/

[7] In addition to those term translation pairs extracted from Wikipedia, it is also helpful to incorporate an existing Japanese-Chinese bilingual lexicon as well as a certain machine translation service between Japanese and Chinese. Those issues are to be examined as a future work.

count.

$$M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) = \\ \Big|\Big\{\, \langle d_J, d_C\rangle \;\Big|\; \big(\, \langle d_J, d_C\rangle \in DD_{JC}(\theta_{JC}) \\ \text{or } \langle d_J, d_C\rangle \in DD_{CJ}(\theta_{JC}) \,\big), \\ P(t_J|d_J) \geq \theta_t,\; P(t_C|d_C) \geq \theta_t \Big\}\Big|$$

Then, we align a Japanese topic t_J to a Chinese topic $t_C (\in TT_C^i)$ which maximizes the count $M_{JC}(t_J, t_C, \theta_t, \theta_{JC})$, only if the count is more than one. Also, we align a Chinese topic t_C to a Japanese topic $t_J (\in TT_J^i)$ which maximizes the count $M_{JC}(t_J, t_C, \theta_t, \theta_{JC})$, only if the count is more than one[8]. For our convenience, we introduce the notations $TA_C(t_J, TT_C^i, \theta_t, \theta_{JC})$ and $TA_J(t_C, TT_J^i, \theta_t, \theta_{JC})$ below in order to denote the results of alignment judgements above:

$$TA_C(t_J, TT_C^i, \theta_t, \theta_{JC}) = \\ \begin{cases} \phi & (\max\limits_{t_C \in TT_C^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) = 1) \\ \operatorname*{argmax}\limits_{t_C \in TT_C^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) & \\ & (\max\limits_{t_C \in TT_C^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) \geq 2) \end{cases}$$

$$TA_J(t_C, TT_J^i, \theta_t, \theta_{JC}) = \\ \begin{cases} \phi & (\max\limits_{t_J \in TT_J^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) = 1) \\ \operatorname*{argmax}\limits_{t_J \in TT_J^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) & \\ & (\max\limits_{t_J \in TT_J^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) \geq 2) \end{cases}$$

As can be obvious from this formalization, the result of cross-lingual topic alignment is not always one-to-one mapping between Japanese and Chinese topics, but it could be one-to-many or many-to-one mapping.

5 Evaluation

5.1 News Articles for Evaluation

As we described in section 3, when extracting Japanese-Chinese term translation pairs from Wikipedia, we collected Japanese and Chinese news articles for the whole one year and extracted candidates of Japanese and Chinese Wikipedia entry titles from them. However, in the evaluation of cross-lingual topic alignment, we used Japanese and Chinese news articles for only one month. This is mainly due to time complexity of the DTM topic modeling toolkit. The DTM topic modeling toolkit performs fairly well even with news articles for only one week. Therefore, in this paper, we report evaluation results with news articles for one month, for which the DTM topic modeling toolkit performs quite well with moderate time complexity.

For the evaluation, we collect Japanese and Chinese news stream text articles during the period from February 25th to March 23rd, 2010. In total, 12,288 Japanese news articles are collected from three newspaper companies Yomiuri, Nikkei, and Asahi, while 22,049 Chinese news articles are collected from People's Daily.

5.2 Evaluation Results

As shown in Table 1, out of the total 30 topics estimated by DTM, we observed 53 bursts (per day) in Japanese topics, out of which we judged 20 bursts as correct, while we observed 40 bursts (per day) in Chinese topics, out of which we judged 17 bursts as correct[9] [10]. Out of those correct 20 bursts in Japanese topics, as shown in Table 2, in Japanese side, the number of bursty topics that are common in Japan and China is only three (per topic) and they are bursty for 6 days in total. One of the three topics is about "Toyota vehicle recalls", while the remaining two are about "Chile earthquake" (in our DTM modeling of Japanese side, the earthquake itself and the Tsunami caused by the earthquake are estimated as two separate topics). In Chinese side, on the other hand, the number of bursty topics that are common in Japan and China is two (per topic) and they are bursty for 4 days in total. One of the two topics is about "Toyota vehicle recalls", while the other one is about "Chile earthquake" (in our DTM modeling of Chinese side, the earthquake itself and the Tsunami caused by the earthquake are not separately estimated).

Next, according to the technique presented in section 4.1, 100 Japanese news articles are cross-lingually aligned to at least one Chinese news article, while 69 Chinese news articles are cross-lingually aligned to at least one Japanese news article. Here, those cross-lingual news article alignment are 100 % correct. Finally, with those cross-lingual news article alignment, Japanese and Chinese bursty topics are 100% correctly aligned according to the technique presented in section 4.2 (6 days / 3 topics from Japanese to Chinese direction and 4 days / 2 topics from Chinese to Japanese direction). In this evaluation, we observed news articles reporting those bursty topics in Tables 1 and 2 more in Japanese than in Chinese. This is why we have article/topic alignment pairs more from J to C than from C to J.

It is also impressive to compare the performance in Tables 1 and 2 with that of without bursty topic detection pre-

[8] Those lower bounds of the number of the pairs of the collected Japanese and Chinese news articles are tuned through a preliminary evaluation.

[9] In this evaluation, we prefer to have more bursts than they actually exist throughout the evaluation period. This is mainly because, even if we have many incorrect bursts in the stage of monolingual burst detection, we can ignore most of those incorrect monolingual bursts in the stage of cross-lingual bursty topic alignment. In the procedure of manual judgement, we judge a burst as "incorrect" when at least one of the following conditions hold: (i) on the day of burst, the set of news articles d which satisfy $P(t|d) \geq \theta_t$ given a topic t do not seem to have consistent content, (ii) even if the set of news articles d which satisfy $P(t|d) \geq \theta_t$ given a topic t seem to have consistent content, such contents can be observed even on the dates other than the burst.

[10] Throughout the evaluation of this paper, manual judgement is performed by one evaluator who is a Chinese native speaker and have a fairly good Japanese proficiency.

Table 1. Evaluation Results: Cross-lingual Alignment of Articles / Bursts / Topics (%)

		country	correct rate (%)	
correct rate of bursty topic (per day)		Japan	37.7	(20/53)
		China	42.5	(17/40)
rate of bursty topics being common in Japan and China (per day)		J to C	30.0	(6/20)
		C to J	23.5	(4/17)
cross-lingual alignment of news articles		J to C	100	(100/100)
		C to J	100	(69/69)
cross-lingual alignment of bursty topics	per day	J to C	100	(6/6)
		C to J	100	(4/4)
	per topic	J to C	100	(3/3)
		C to J	100	(2/2)

Table 2. Bursty Topics and # of Pairs of Japanese / Chinese News Articles on Each Day

topics both in Japanese and Chinese			Dates					
			February, 2010				March, 2010	
			25	26	27	28	1	2∼23
Toyota vehicle recalls	Japan		burst (36 pairs)					no bursty topics common in Japan and China
	China		burst (18 pairs)					
Chile earthquake	Japan	the earthquake			burst (21 pairs)	burst (9 pairs)		
		Tsunami			burst (6 pairs)	burst (22 pairs)	burst (6 pairs)	
	China				burst (17 pairs)	burst (28 pairs)	burst (6 pairs)	
Evaluation Results of Topic Alignment: J to C 100% (3/3) C to J 100% (2/2)								

sented in our previous work [8]. With the same data set as in our previous work [8], correct rates of cross-lingual alignment of news articles improve from 53∼63% to 100%, while those of cross-lingual topic alignment improve from 67∼80% to 100%, simply by detecting bursty topics in advance of cross-lingual topic alignment and aligning topics that are bursty for both Japanese and Chinese. In our previous work [8], topics that are cross-lingually aligned without bursty topic detection include those on Japanese and Chinese domestic economies, which are observed throughout the whole period of the data set and are judged as non-bursty. Our reference cross-lingually aligned topics between Japanese and Chinese do not include those on Japanese and Chinese domestic economies, and hence they damaged the correct rate of cross-lingual topic alignment in our previous work [8].

6 Related Work

Wang et al. [13] studied how to detect correlated bursty topic patterns across multiple text streams such as multilingual news streams, where their method concentrated on detecting correlated bursty topic patterns based on the similarity of temporal distribution of tokens. Boyd-Graber and Blei [4], De Smet and Moens [5], and Zhang et al. [15] concentrated on applying variants of topic models which have certain functions of bridging cross-lingual gaps by exploiting clues such as translation knowledge from bilingual lexicon or distribution of named entities. Compared with those previous works, the approach we take in this paper is different in that we focus on a time series topic model and align time series topics across two languages. It is one of our future works to introduce those other models and compare them with our proposed framework in terms of effectiveness of aligning time series topics across two languages.

7 Concluding Remarks

This paper studied issues regarding topic modeling of information flow in multilingual news streams. To the results of estimating distribution of topics in Japanese / Chinese news streams, we applied Kleinberg's modeling of bursts [9], and detected bursty topics for both Japanese and Chinese news. We showed that, by aligning topics that are bursty for both Japanese and Chinese, correct rate of cross-lingual topic alignment of about 67∼80% when without bursty topic detection [8] improved up to 100%. Future works include precise evaluation of recall[11], where we an-

[11] In terms of recall and precision, the proposed method achieved 100% precision, while the evaluation of recall is one of our future works.

notate topic alignment information to certain random samples of Japanese and Chinese time series news stream, and then, examine whether they are actually detected by the proposed method. Another future work is aligning different opinions of a same topic between two countries, where we suppose that, with our topic modeling framework, each of different opinions is allocated to a distinct topic of DTM topic modeling, and can be easily cross-lingually aligned between two languages.

References

[1] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proc. ICWSM*, pages 19–26, 2008.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proc. 23rd ICML*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In *Proc. 25th UAI*, pages 75–82, 2009.

[5] W. De Smet and M.-F. Moens. Cross-language linking of news stories on the Web using interlingual topic modelling. In *Proc. 2nd SWSM*, pages 57–64, 2009.

[6] D. K. Evans, L.-W. Ku, Y. Seki, H.-H. Chen, and N. Kando. Opinion Analysis across Languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task. In *Proc. 3rd CLIP*, pages 456–463, 2007.

[7] T. Fukuhara, T. Utsuro, and H. Nakagawa. Cross-lingual concern analysis from multilingual weblog articles. In *Proc. 6th Inter. Workshop on Social Intelligence Design*, pages 55–64, 2007.

[8] S. Hu, Y. Takahashi, L. Zheng, T. Utsuro, M. Yoshioka, N. Kando, T. Fukuhara, H. Nakagawa, and Y. Kiyota. Cross-lingual topic alignment in time series Japanese / Chinese news. In *Proc. 26th PACLIC*, pages 532–541, 2012.

[9] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th SIGKDD*, pages 91–101, 2002.

[10] D. Koike, Y. Takahashi, T. Utsuro, M. Yoshioka, and N. Kando. Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proc. 6th IJCNLP*, 2013.

[11] H. Nakasaki, M. Kawaba, S. Yamazaki, T. Utsuro, and T. Fukuhara. Visualizing cross-lingual/cross-cultural differences in concerns in multilingual blogs. In *Proc. ICWSM*, pages 270–273, 2009.

[12] Y. Takahashi, T. Utsuro, M. Yoshioka, N. Kando, T. Fukuhara, H. Nakagawa, and Y. Kiyota. Applying a burst model to detect bursty topics in a topic model. In *JapTAL 2012*, volume 7614 of *LNCS*, pages 239–249. Springer, 2012.

[13] X. Wang, C. Zhai, and R. S. X. Hu. Mining correlated bursty topic patterns from coordinated text streams. In *Proc. 13th SIGKDD*, pages 784–793, 2007.

[14] R. Yangarber, C. Best, P. von Etter, F. Fuart, D. Horby, and R. Steinberger. Combining information about epidemic threats from multiple sources. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pages 41–48, 2007.

[15] D. Zhang, Q. Mei, and C.-X. Zhai. Cross-lingual latent topic extraction. In *Proc. 48th ACL*, pages 1128–1137, 2010.