Cross-Lingual Topic Alignment in Time Series Japanese / Chinese News

Shuo Hu **Yusuke Takahashi** **Liyi Zheng** **Takehito Utsuro**
Graduate School of Systems and Information Engineering, University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Masaharu Yoshioka
Graduate School of Information
Science and Technology,
Hokkaido University,
Sapporo, 060-0808, Japan

Noriko Kando
National Institute
of Informatics,
Tokyo 101-8430, Japan

Tomohiro Fukuhara
National Institute of Advanced
Industrial Science and Technology
Tsukuba, 305-8568 Japan

Hiroshi Nakagawa
Information Technology Center,
University of Tokyo, Tokyo 113-0033, Japan

Yoji Kiyota
NEXT Co., Ltd.,
Tokyo, 108-0075, Japan

Abstract

Among various types of recent information explosion, that in news stream is also a kind of serious problems. 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. Especially, as a topic model, this paper employs DTM, but not LDA, since it can consider correspondence between topics of consecutive dates. 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.

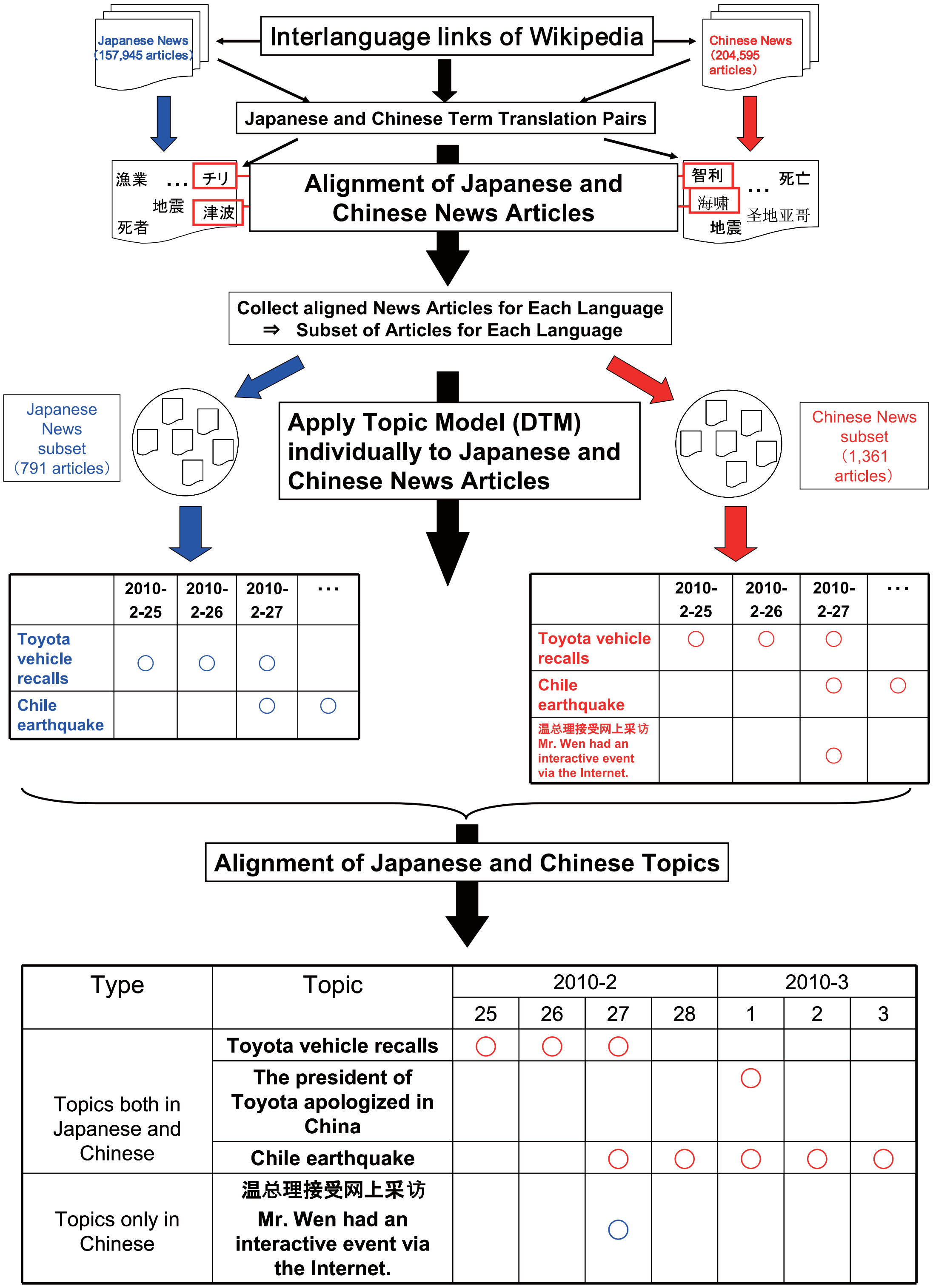
1 Introduction

Among various types of recent information explosion, that in news stream is also a kind of serious problems. 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) (Blei et al., 2003) and DTM (dynamic topic model) (Blei and Lafferty, 2006) 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. 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 first cross-lingually align Japanese and Chinese news articles. Then, after collecting those cross-lingually aligned news article pairs, we then apply DTM to those collected news articles and estimate time series monolingual topic models for both Japanese and Chinese. Finally, those monolingual

26th Pacific Asia Conference on Language,Information and Computation pages 498–507


Interlanguage links of Wikipedia
Japanese News（157,945 articles）
Chinese News（204,595 articles）
Japanese and Chinese Term Translation Pairs
漁業 … チリ
地震 津波
死者
智利 … 死亡
海啸 圣地亚哥
地震
Alignment of Japanese and Chinese News Articles
Collect aligned News Articles for Each Language
⇒ Subset of Articles for Each Language
Japanese News subset（791 articles）
Apply Topic Model (DTM) individually to Japanese and Chinese News Articles
Chinese News subset（1,361 articles）
Alignment of Japanese and Chinese Topics

	2010-2-25	2010-2-26	2010-2-27	…
Toyota vehicle recalls	○	○	○	
Chile earthquake			○	○

	2010-2-25	2010-2-26	2010-2-27	…
Toyota vehicle recalls	○	○	○	
Chile earthquake			○	○
温总理接受网上采访 Mr. Wen had an interactive event via the Internet.			○	

Type	Topic	2010-2				2010-3		
		25	26	27	28	1	2	3
Topics both in Japanese and Chinese	Toyota vehicle recalls	○	○	○				
	The president of Toyota apologized in China					○		
	Chile earthquake			○	○	○	○	○
Topics only in Chinese	温总理接受网上采访 Mr. Wen had an interactive event via the Internet.			○				

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

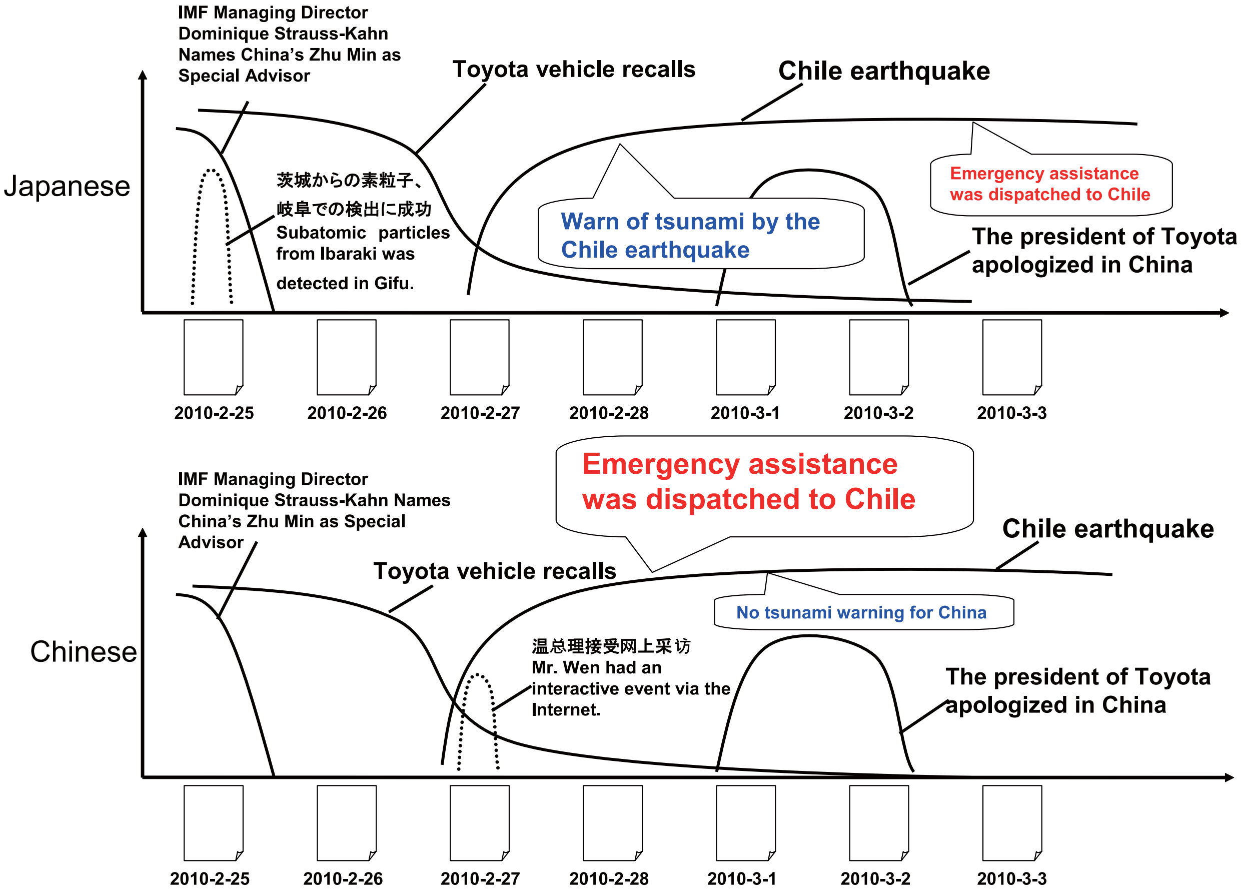

IMF Managing Director Dominique Strauss-Kahn Names China's Zhu Min as Special Advisor
Toyota vehicle recalls
Chile earthquake
Japanese
茨城からの素粒子、
岐阜での検出に成功
Subatomic particles from Ibaraki was detected in Gifu.
Warn of tsunami by the Chile earthquake
Emergency assistance was dispatched to Chile
The president of Toyota apologized in China
2010-2-25
2010-2-26
2010-2-27
2010-2-28
2010-3-1
2010-3-2
2010-3-3
IMF Managing Director Dominique Strauss-Kahn Names China's Zhu Min as Special Advisor
Emergency assistance was dispatched to Chile
Chile earthquake
Toyota vehicle recalls
No tsunami warning for China
Chinese
温总理接受网上采访
Mr. Wen had an interactive event via the Internet.
The president of Toyota apologized in China
2010-2-25
2010-2-26
2010-2-27
2010-2-28
2010-3-1
2010-3-2
2010-3-3

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

topics are cross-lingually aligned considering cross-lingual alignment of Japanese and Chinese news articles.

Figure 2 shows an example of estimating time series 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 Topic Model

As a time series topic model, this paper employs DTM (dynamic topic model) (Blei and Lafferty, 2006). Unlike LDA (Latent Dirichlet Allocation) (Blei et al., 2003), 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 $(n = 1, \ldots, K)$ as well as that $p(z_n \mid d)$ $(n = 1, \ldots, K)$ 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 $p(z_n \mid d)$ $(n = 1, \ldots, K)$ by a Blei's toolkit[1], where for the number of topics $K = 10$, as well as $\alpha = 0.01$.

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. 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. Figure 3 describes the rough idea of how to extract translation pairs of Japanese terms and simplified Chinese character terms from interlanguage links of Wikipedia.

Let a Japanese Wikipedia entry e_J to be denoted as $e_J = \langle J_0, \{J_r^1, \ldots, J_r^l\}\rangle$, where J_0 is the title of the entry e_J, and $J_r^1, \ldots, J_r^l$ are redirects of the entry e_J. Let e_C be a Chinese Wikipedia entry for which at least one of a interlanguage link from e_J to e_C or that from e_C to e_J exists. In the Chinese version of Wikipedia, entries including entry titles are usually written in traditional Chinese characters and equivalent terms in simplified Chinese characters are listed as redirects of terms in traditional Chinese characters. Thus, e_C is denoted as $e_C = \langle T_0, \{S_r^1, \ldots, S_r^k, T_r^{k+1}, \ldots, T_r^h\}\rangle$, where T_0 is the title string of the entry e_C in traditional Chinese characters, $S_r^1, \ldots, S_r^k$ are redirects of the entry e_C in simplified Chinese characters, and $T_r^{k+1}, \ldots, T_r^h$ are redirects of the entry e_C in traditional Chinese characters.

Since it is not easy for us to automatically distinguish character codes for simplified Chinese and traditional Chinese, we utilize news articles of the collection of one year that are written in simplified Chinese characters, and employ the following procedure to extract translation pairs of Japanese terms and simplified Chinese character terms. First, suppose that we detect one of those redirects of the entry e_C in simplified Chinese characters, namely S_r^i, in a Chinese news article written in simplified Chinese characters. Then, following the interlanguage link between the entries e_C and e_J, we collect the term translation pairs below between Japanese and simplified Chinese characters into the set $JS(\langle e_J, e_C, S_r^i\rangle)$ of term translation pairs including S_r^i:

$$\begin{aligned} JS(\langle e_J, e_C, S_r^i\rangle) \;\; = \;\; & \big\{\langle J_0, S_r^i\rangle,\ \langle J_r^1, S_r^i\rangle, \ldots, \\ & \langle J_r^l, S_r^i\rangle\big\} \end{aligned}$$

Then, we collect the term translation pairs in the whole sets $JS(\langle e_J, e_C, S\rangle)$ into JS_W:

$$JS_W \;\; = \bigcup_{\langle e_J, e_C, S\rangle} JS(\langle e_J, e_C, S\rangle)$$

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[2], Nikkei[3], and Asahi[4], while 204,595 Chinese news articles are collected from People's Daily[5]. 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[6].

4 Cross-lingual Topic Alignment

This section proposes the whole framework of cross-lingual topic alignment, where its major steps are illustrated in the overall flow in Figure 1.

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

[2] `http://www.yomiuri.co.jp/`

[3] `http://www.nikkei.com/`

[4] `http://www.asahi.com/`

[5] `http://www.people.com.cn/`

[6] 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.

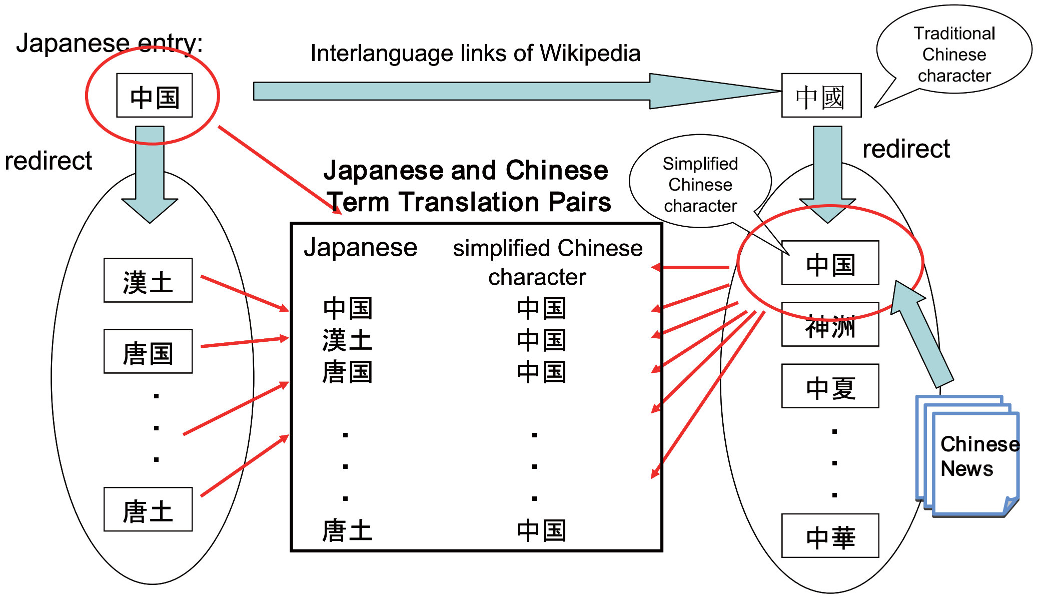

Japanese entry:
中国
Interlanguage links of Wikipedia
中國
Traditional Chinese character
redirect
redirect
Japanese and Chinese Term Translation Pairs
Japanese
simplified Chinese character
中国
中国
漢土
中国
唐国
中国
唐土
中国
漢土
唐国
唐土
Simplified Chinese character
中国
神洲
中夏
中華
Chinese News

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

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 10).

More specifically, 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 included in JS_W, which are shared between d_J and d_C:

$$N_{JC}(d_J, d_C) = \Big|\{\langle J, S\rangle (\in JS_W) \mid J \text{ appears in } d_J. \; S \text{ appears in } d_C.\}\Big|$$

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:

$$DD_{JC}(\theta_{JC}) = \Big\{\langle d_J, d_C\rangle \mid N_{JC}(d_J, d_C) \geq \theta_{JC}, \; d_C = \operatorname*{argmax}_{d'_C} N_{JC}(d_J, d'_C)\Big\}$$

$$DD_{CJ}(\theta_{JC}) = \Big\{\langle d_J, d_C\rangle \mid N_{JC}(d_J, d_C) \geq \theta_{JC}, \; d_J = \operatorname*{argmax}_{d'_J} N_{JC}(d'_J, d_C)\Big\}$$

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}.

4.2 Cross-Lingual Alignment of Topics

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

First, for each date, all the Japanese news articles are collected from the sets $DD_{JC}(\theta_{JC})$ and $DD_{CJ}(\theta_{JC})$. Next, collected Japanese news articles are accumulated during the period of evaluation, and the DTM topic modeling toolkit is applied to the accumulated news articles and K topics are estimated for each date during the period of evaluation. Then, on the i-th day of the period of evaluation, we have the set TT^i_J of estimated Japanese topics.

In the similar way, all the Chinese news articles are collected from the sets $DD_{JC}(\theta_{JC})$ and $DD_{CJ}(\theta_{JC})$. Collected Chinese news articles are accumulated during the period of evaluation, and the DTM topic modeling toolkit is applied to the accumulated news articles and K topics are estimated for each date during the period of evaluation. Then, on the i-th day of the period of evaluation, we have the set TT_C^i of estimated Chinese topics.

Once we have the sets TT_J^i and TT_C^i on the i-th day, we align the Japanese and Chinese topics of TT_J^i and TT_C^i according to the following procedure. First, for each Japanese topic $t_J(\in TT_J^i)$, we collect news articles d_J which satisfy $P(t_J|d_J) \geq \theta_t$ (in this paper, θ_t is 0.6). In the similar way, for each Chinese topic $t_C(\in TT_C^i)$, we collect news articles d_C which satisfy $P(t_C|d_C) \geq \theta_t$. Then, out of the pairs of collected news articles $\langle d_J, d_C\rangle$, we count the number of those 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 count.

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

Finally, 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. 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:

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

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

4.3 Cross-Lingual Alignment of Time Series Topic Sequence

Suppose that the period of evaluation consists of n consecutive days, then the procedure of cross-lingual alignment of time series topic sequence is described as follows.

First, let $Q_J = TT_J^1, TT_J^2, \ldots, TT_J^n$ be the sequence of sets of Japanese topics, which are estimated through the DTM topic modeling toolkit, and each set TT_J^i of topics is for the news articles published on the i-th day of the evaluation period. Also, let $Q_C = TT_C^1, TT_C^2, \ldots, TT_J^n$ be the sequence of sets of Chinese topics, which are estimated through the DTM topic modeling toolkit. Then, for each of the n consecutive days, cross-lingual topic alignment is performed according to the following procedure:[7]

- On the i-th day, for each Japanese topic $t_J(\in TT_J^i)$, obtain the topic alignment judgement result $TA_C(t_J, TT_C^i, \theta_t, \theta_{JC})$.

- Similarly on the i-th day, for each Chinese topic $t_C(\in TT_C^i)$, obtain the topic alignment judgement result $TA_J(t_C, TT_J^i, \theta_t, \theta_{JC})$.

[7] In DTM, on the i-th day, it is possible to refer to topic models of neighboring days such as $i-1$-th and $i+1$-th days. Although in our cross-lingual topic alignment technique, we do not utilize such information, the evaluation results of cross-lingual topic alignment did not conflict with those of topics of neighboring days.

Table 1: Evaluation Results (Correct Rate): Alignment of Japanese / Chinese News Articles (%)

(a) *With* News Articles on Japanese / Chinese Domestic Economy

Date	Japanese to Chinese		Chinese to Japanese	
February 25, 2010	53.0	(26/49)	54.8	(40/73)
February 26, 2010	62.1	(18/29)	62.5	(15/24)
February 27, 2010	76.7	(23/30)	88.6	(31/35)
February 28, 2010	88.2	(30/34)	87.8	(36/41)
March 1, 2010	58.7	(27/46)	54.7	(35/64)
March 2, 2010	43.5	(10/23)	40.0	(12/30)
March 3, 2010	61.1	(22/36)	25.8	(25/97)
Total	63.2	(156/247)	53.3	(194/364)

(b) *Without* News Articles on Japanese / Chinese Domestic Economy

Date	Japanese to Chinese		Chinese to Japanese	
February 25, 2010	83.9	(26/31)	93.0	(40/43)
February 26, 2010	94.7	(18/19)	100	(15/15)
February 27, 2010	76.7	(23/30)	88.6	(31/35)
February 28, 2010	88.2	(30/34)	87.8	(36/41)
March 1, 2010	93.1	(27/29)	87.5	(35/40)
March 2, 2010	90.1	(10/11)	92.3	(12/13)
March 3, 2010	95.7	(22/23)	67.6	(25/37)
Total	88.1	(156/177)	86.6	(194/224)

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 first 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 Cross-Lingual Alignment of News Articles

After we cross-lingually align Japanese and Chinese news articles by the method we presented in section 4.1, each of 791 Japanese articles is aligned to a Chinese news article, while each of 1,361 Chinese articles is aligned to a Japanese news article. Out of evaluation results for the whole one month, Table 1 shows the excerpts for that of one week (February 25th to March 3rd, 2010). Table 1 (a) shows the results without manually removing a certain subset of news articles, where correct rate of cross-lingual alignment of news articles is around 60% on the average. Relatively low correct rate is mainly due to Japanese and Chinese news articles on domestic

Table 2: Evaluation Results: Cross-Lingual Topic Alignment (*with* news articles on Japanese / Chinese domestic economy)

		Dates							
		February, 2010				March, 2010			
Type	Topic	25	26	27	28	1	2	3	4 ∼ 23
topics both in Japanese and Chinese (correct alignment)	Toyota vehicle recalls	correct topic alignment							topics are not cross-lingually aligned.
	The president of Toyota apologized in China					correct topic alignment			
	Chile earthquake			correct topic alignment					
	IMF Managing Director Dominique Strauss-Kahn Names China′s Zhu Min as Special Advisor	correct topic alignment							
Topics only in Chinese	Mr. Wen had an interactive event via the Internet			alignment error					
topics both in Japanese and Chinese (alignment error)	Domestic Economy News	topics are aligned every day with error.							
Evaluation results (correct rate) :		Japanese to Chinese	80.0% (4/5)			Chinese to Japanese			66.7% (4/6)

economies. Both Japanese and Chinese news articles on domestic economies include numerical figures as well as technical terms on the economy domain, although their contents are not cross-lingually related to each other at all. It is also interesting to note that February 27th and 28th, 2010 were Saturday and Sunday. It is quite natural that, since much less news articles on domestic economies are published on holidays, correct rates on those dates are apparently higher than those on other dates.

Next, we manually remove those news articles on domestic economies, and measure the correct rates of cross-lingual alignment of news articles as we show in Table 1 (b). In this case, correct rates drastically go up to more than 85% on the average. One obvious future plan for automatically removing news articles on domestic economies for both languages is to simply apply a well studied techniques of burst detection such as the one proposed in Kleinberg (2002). Since, both in Japanese and in Chinese, news articles on domestic economies are constantly published on every week day, it is strongly estimated that they are not detected at all.

5.3 Cross-Lingual Alignment of Topics

Next, the DTM topic modeling toolkit is applied to the 791 Japanese articles as well as the 1,361 Chinese articles introduced in the previous section. Then, cross-lingual topic alignment procedure presented in section 4.2 is applied to them[8], whose evaluation results are shown in Table 2.

Out of the evaluation period of the whole one month, cross-lingually aligned topics are detected only for the first one week, except that the topics on domestic economies are cross-lingually aligned every day throughout the whole one month. Among the remaining five topic alignment results, only the one “Mr. Wen had an interactive event via the In-

[8] When applying the cross-lingual topic alignment procedure, we keep errors in the process of cross-lingual alignment of news articles, which means that only about 50∼60% of the results of cross-lingual alignment of news articles are correct.

ternet" is alignment error. This topic is somehow concerned with a Chinese domestic issue and the topic itself is successfully estimated only in Chinese. About ten Chinese news articles are aligned to exactly the same Japanese article and this cross-lingual article alignment result causes the cross-lingual topic alignment error. Considering those evaluation results, if we count a sequence of cross-lingual topic alignment on consecutive days as one if aligned topics on those consecutive days are exactly the same, the correct rate of Japanese to Chinese topic alignment is 80.0%, while that of Chinese to Japanese direction is 66.7%.

In this evaluation result, one erroneous topic alignment from Japanese to Chinese and one of the two erroneous topic alignments from Chinese to Japanese are the ones about domestic economies. Thus, if we remove those erroneous alignment results of topics on domestic economies, we have the correct rate of Japanese to Chinese topic alignment as 100% (=5/5) and that of Chinese to Japanese direction as 80.0% (=4/5).

6 Related Works

Wang et al. (2007) 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. Unlike the method of Wang et al. (2007), in this paper, we do not utilize burst detection techniques, but employ a time series topic model and cross-lingually align time series topics utilizing translation knowledge automatically extracted from Wikipedia.

Boyd-Graber and Blei (2009), De Smet and Moens (2009), Zhang et al. (2010), and Jagarlamudi and Daumé III (2010) 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 studies issues regarding topic modeling of information flow in multilingual news streams. Based on the results of estimating distribution of topics in Japanese / Chinese news streams, this paper proposed how to analyze cross-lingual alignment of topics in time series Japanese / Chinese news streams. Evaluation results show that the proposed method is quite effective in discovering cross-lingual topic alignment between Japanese and Chinese news streams.

Future works include precise evaluation of recall, where we annotate 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. Also, we plan to incorporate our recently invented technique (Takahashi et al., 2012) which is capable of detecting bursty topics within a time series text stream, and then cross-lingually align Japanese and Chinese bursty topics.

References

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

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

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

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

J. Jagarlamudi and H. Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Proc. 32nd ECIR*, pages 444–456.

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

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

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

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