日中時系列ニュースにおける バースト・トピックの推定と二言語間対応付け*

胡 碩† 高橋 佑介† 鄭 立儀† 宇津呂 武仁‡ 吉岡 真治§ 神門 典子¶
筑波大学大学院 システム情報工学研究科†
筑波大学 システム情報系‡ 北海道大学大学院 情報科学研究科§ 国立情報学研究所¶

1 はじめに

現代の情報社会においては，多種多様な情報が氾濫し，いわゆる情報爆発の問題が深刻であり，氾濫する情報の集約や，俯瞰を行うための技術の確立が強く望まれている．中でも，情報爆発が最も顕著に現れているのはウェブであり，ウェブ上の情報爆発の問題に取り組んだ研究が盛んに行われている．例えば，バースト解析の技術においては，ストリームデータの時間軸方向の密度から世の中の異変や特異な出来事を捉えることができる．また，別のアプローチとして，トピックモデルのように文書集合における主要なトピックを推定することのできる技術も存在する．

ここで，バースト解析は，一般には，電子メールやウェブ上のニュース記事のようなストリームデータに対して適用される．そこでは，ある時からある話題に関する記述が急激に増加するような現象が起こることがあり，こういった現象を，ある話題に関するバーストと呼ぶ．代表的なアルゴリズムである Kleinberg のバースト解析 [4] では，時系列に沿った各キーワードのバースト度の変化や，バーストしているか否かの判定，バースト度によるキーワードのランク付けをすることができる．

一方，トピックモデルにおいては，文書が生成される背景には，潜在的にいくつかのトピックがあることを想定し，文書の生成尤度を高めるようにモデルのパラメータを訓練する．トピックモデルの一種である DTM (dynamic topic model) [1] においては，時系列情報を持つ文書集合を情報源として，時系列にそって，各単位時間ごとに，文書ごとのトピックの分布と，トピックごとの語の分布を求めることができる．

以上をふまえて，本論文では，日本語および中国語の二言語の時系列ニュースを対象として，DTM によってトピックのバースト解析を行う．そして，日中両方でバーストした時系列トピックに対して，日中間でトピックの対応をとる手法を提案する．本研究により日中間でトピックの対応を推定した事例を図 1 に示す．本研究では，このように，日中の時系列ニュースにおけるバースト・トピックの二言語間対応を分析することにより，日中間の関心や意見の差異を検出するための基盤技術を確立する．我々は，文献 [3] において，トピックのバーストを考慮しない場合に，日中間のニュース記事対応付け精度が約 50∼60%，日中間のトピック対応付け精度が約 65∼80%となることを示したが，本論文において，トピックのバーストを考慮することにより，これらの精度が改善することを示す．

2 時系列トピックモデルのバースト解析

2.1 トピックモデル

本研究では，トピックモデルとして DTM (dynamic topic model) [1] を用いる．DTM は，語 w の列によって表現される時間情報を含んだ文書の集合と，トピック数 K を入力とし，各単位時間について，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $p(w|z_n)$ $(w \in V)$，及び，各文書 b におけるトピック z_n の確率分布 $p(z_n|b)$ $(n = 1, \ldots, K)$ を推定する．ここで，V は文書中に出現する語の集合である．

DTM は，潜在的ディリクレ配分法 (LDA, Latent Dirichlet Allocation) [2] とは異なり，文書集合中の時系列情報を考慮しているため，日付等の単位時間を超えて同一トピックを追跡可能である．

*Estimation and Cross-Lingual Alignment of Bursty Topics in Time Series Japanese / Chinese News Streams

†Shuo Hu, Yusuke Takahashi, LiYi Zheng, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba
§Masaharu Yoshioka, Graduate School of Information Scienceand Technology, Hokkaido University
¶Noriko Kando, National Institute of Informatics

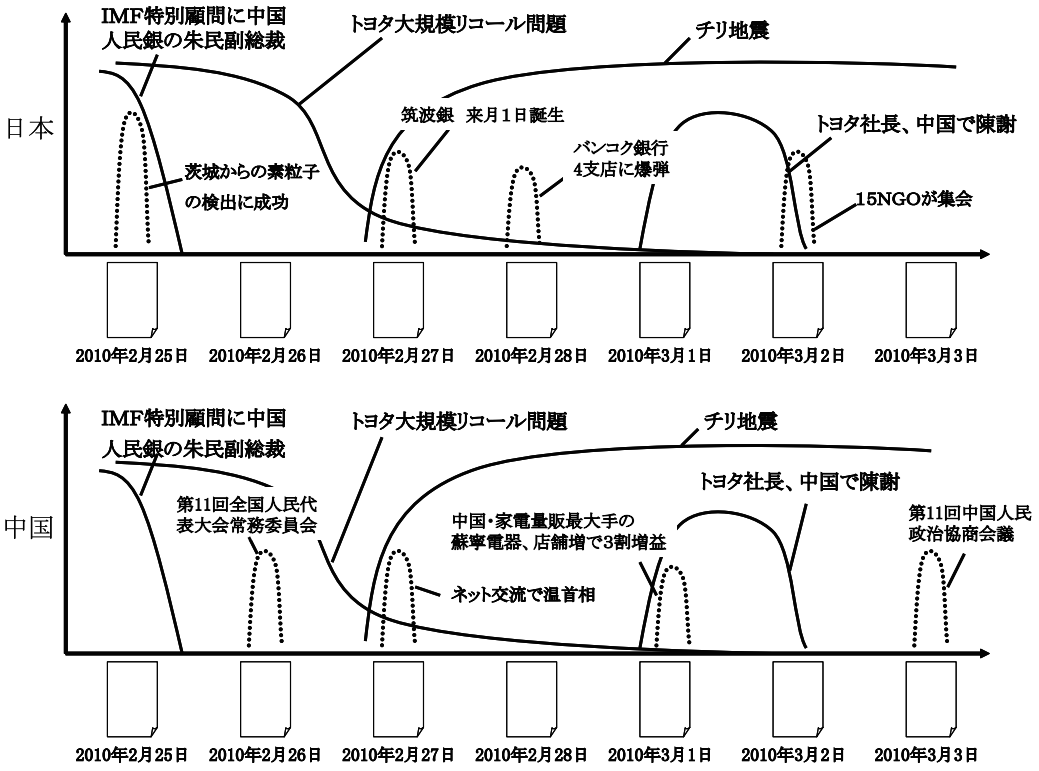

図 1: 日中時系列ニュースにおけるトピックの分析

本論文では，$p(w|z_n)$ $(w \in V)$，及び，$p(z_n|b)$ $(n = 1, \ldots, K)$ の推定においては，Blei らによって公開されたツール[1]を用いた．ハイパーパラメータ α と，トピック数 K は，それぞれ $\alpha = 0.01$，$K = 30$ とした．

本研究では，一日ごとに，各トピックに対してニュース記事を一対一で割り当てることで，トピックごとのニュース記事集合の要素数を測ることとした．

ある日における文書集合を D，トピック数を K，一つの文書を $d(d \in D)$ とすると，トピック $z_n(n = 1, \ldots, K)$ のニュース記事集合 $D(z_n)$ は以下の式で表される．

$$D(z_n) = \left\{ d \in D \middle| z_n = \underset{z_u (u=1,\ldots,K)}{\operatorname{argmax}} p(z_u|d) \right\}$$

これはつまり，文書 d におけるトピックの分布において，文書 d に確率が最大のトピックを割り当てていることになる．

2.2 トピックモデルのバースト解析

Kleinberg のバースト解析 [4] は，各日における文書数 d_t と，その日の関連文書数 r_t を入力として，解析期間におけるバースト状態と非バースト状態を切り分けて出力する手法である．したがって，Kleinberg の手法を用いてトピックのバーストを測るためには，各日における各トピックの関連文書数 r_t が得られれば良い．そこで，本手法ではトピック z_n の関連文書数 r_t を以下のように定義することで，トピックのバースト解析 [6] を行う．

$$r_t = \sum_b p(z_n|b)$$

これより，解析期間における全ての関連記事数 $R = \sum_{t=1}^{m} r_t$ が求まり，それを解析期間における全ての記事の数 $D = \sum_{t=1}^{m} d_t$ で割ることにより，解析期間全体における期待値 $p_0 = R/D$ を算出する．ただし，バースト状態の期待値 p_1 と非バースト状態の期待値 p_0 の間の係数 $s(p_1 = p_0 s)$ としては，$s = 2.8$ の値を用いる．また，状態遷移を妨げるためのパラメータ γ としては，$\gamma = 1$ の値を用いる．

3 トピックの二言語間対応の推定

3.1 ニュース記事の日中対応の推定

ニュース記事の日中対応推定においては，まず，2.2 節の手法により，バーストするトピックにおける記

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

事の中で，$P(t|d) \geq \theta_t$ ($\theta_t = 0.6$) という条件を満たす日中記事を集めて，日本語および中国語のニュース記事集合を作る．そして，一日の単位で，日本語ニュース記事と中国語ニュース記事との間で，共有する日中対訳語組数[2]を求めて，一定の値 θ_{JC} (本稿では，θ_{JC} を 8 に設定した) 以上の共有日中対訳語組を持つ日本語ニュース記事と中国語ニュース記事の組に対して，ニュース記事の日中対応を推定する．日中ニュース記事の間で，共有する対訳組の集合の大きさ $N_{JC}(d_J, d_C)$ を以下のように定義する．

$$N_{JC}(d_J, d_C) = \left|\{\langle J, S\rangle (\in JS_W) \mid J \text{ は } d_J \text{中に出現する．} S \text{ は } d_C \text{中に出現する．}\}\right|$$

ここで，d_J は日本語ニュース記事である．d_C は中国語ニュース記事である．θ_{JC} 以上の共有日中対訳語組数を持つ日中ニュース記事組の集合 $DD_{JC}(\theta_{JC})$ および $DD_{CJ}(\theta_{JC})$ を以下の式で定義する．

$$DD_{JC}(\theta_{JC}) = \left\{\langle d_J, d_C\rangle \mid N_{JC}(d_J, d_C) \geq \theta_{JC}, d_C = \underset{d'_C}{\operatorname{argmax}} N_{JC}(d_J, d'_C)\right\}$$

$$DD_{CJ}(\theta_{JC}) = \left\{\langle d_J, d_C\rangle \mid N_{JC}(d_J, d_C) \geq \theta_{JC}, d_J = \underset{d'_J}{\operatorname{argmax}} N_{JC}(d'_J, d_C)\right\}$$

3.2 トピックの日中対応の推定

本節では，前節で作成した日中ニュース記事組の集合 DD_{JC} もしくは DD_{CJ} 中に含まれる日本語記事 d_J および d_C のみを対象として，以下の手順により，トピックの日中対応の推定を行う．

まず，$DD_{JC}(\theta_{JC})$ または $DD_{CJ}(\theta_{JC})$ に含まれる記事組を抽出し，その要素数を $M_{JC}(t_J, t_C, \theta_t, \theta_{JC})$ とする．そして，以下の式により，日本語トピック集合 TT_J^i 中のトピックのうち，中国語トピック t_C に対応するものを同定する．同様に，中国語トピック集合 TT_C^i 中のトピックのうち，日本語トピック t_J に対応するものを同定する．

$$TA_J(t_C, TT_J^i, \theta_t, \theta_{JC}) = \begin{cases} \text{出力なし} & (\max_{t_J \in TT_J^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) \leq 1) \\ \underset{t_J \in TT_J^i}{\operatorname{argmax}} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) & (\max_{t_J \in TT_J^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) \geq 2) \end{cases}$$

$$TA_C(t_J, TT_C^i, \theta_t, \theta_{JC}) = \begin{cases} \text{出力なし} & (\max_{t_C \in TT_C^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) \leq 1) \\ \underset{t_C \in TT_C^i}{\operatorname{argmax}} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) & (\max_{t_C \in TT_C^i} M_{JC}(t_J, t_C, \theta_t, \theta_{JC}) \geq 2) \end{cases}$$

4 分析および評価

トピックモデルのバースト解析は，2010 年 2 月 25 日から，3 月 23 日までの一ヶ月間の読売新聞[3]，日経新聞[4]，および，朝日新聞[5]の三紙の 12,288 ニュース記事，および，人民日報[6]の 22,049 ニュース記事を対象として行った．この一ヶ月において，トピックモデル DTM により，30 個のトピックの中で，日本側のバースト日数は 53 日間，中国側のバースト日数は 40 日間であった．そのうち，バーストが適切であった日数は，日本側は 20 日間，中国側は 17 日間であった．適切なバースト・トピックにおいて，日中間で対応する日数は，日本側は 6 トピック × 日，中国側は 4 トピック × 日であった．そして，3.1 節の手法により，バーストしたトピックに所属する記事集合において，日中二言

表 1: バースト，記事およびトピックの日中対応の評価結果 (%)

	国	精度 (%)	
バースト・トピックの日単位の適合率	日本	37.7	(20/53)
	中国	42.5	(17/40)
バースト・トピックが日中で対応する割合 (日単位)	日本	30.0	(6/20)
	中国	23.5	(4/17)
ニュース記事の日中対応精度	日本	100	(100/100)
	中国	100	(69/69)
トピックの日中対応精度	日本	100	(2/2)
	中国	100	(2/2)

[2]中国語の簡体字ニュース記事中に出現する簡体字キーワードに対して，Wikipedia の言語間リンクを用いてエントリの日中対応を抽出し，日中対訳語組の集合を作成する．ただし，中国語版 Wikipedia においては，簡体字キーワードは，繁体字のエントリタイトルを持つエントリ中のリダイレクトとして登録されているため，これを利用する．2009 年 6 月 1 日 ~2010 年 5 月 31 日の 1 年分の 157,945 日本語ニュース記事，および，204,595 中国語ニュース記事を利用することにより，78,519 組の日中対訳語組を抽出した．

[3]http://www.yomiuri.co.jp/
[4]http://www.nikkei.com/
[5]http://www.asahi.com/
[6]http://www.people.com.cn/

表 2: 各日のトピックにおけるバーストの有無および日中対応記事組数

				2010 年 2 月				2010 年 3 月	
				25 日	26 日	27 日	28 日	1 日	2 日〜23 日
日中共通	トヨタリコール問題	日本		有 (36 組)					日中共通のバーストなし
		中国		有 (18 組)					
	チリ地震	日本	チリ地震			有 (21 組)	有 (9 組)		
			チリ地震による津波			有 (6 組)	有 (22 組)	有 (6 組)	
		中国				有 (17 組)	有 (28 組)	有 (6 組)	
中国のみ	砂嵐								省略
	温首相がネットで交流					有			
	両会について								
	米ロ核条約								
	アフガンテロ								
	国際消費者権益日								
日本のみ	バンクーバーオリンピック			有	有	有	有	有	省略
	イラク選挙								
	イラク議会選								
	ワシントン条約								
	タクシン派抗議								
	水俣病について								
	アカデミー賞								
	小林議員辞職								
トピック単位での日中対応の評価結果: 日本語トピック 100.0%(2/2) 中国語トピック 100.0%(2/2)									

語間で対応が付けられた日本語ニュース記事数は 100 件，中国語ニュース記事数は 69 件であった．

この結果において，日本語から中国語，および，中国語から日本語へのニュース記事対応付け精度は，いずれも 100%であった．次に，日中間で対応が付けられた日本語ニュース記事 100 記事，中国語ニュース記事 69 記事を対象として，日中間でトピックの対応の推定を行った．この評価結果を表 1 に示す．また，各日のトピックにおけるバーストの有無および日中対応記事組数を表 2 に示す．

5 関連研究

文献 [7] においては，複数の時系列テキストを対象として，相互に関連するバースト・トピックを検出する手法を提案している．具体的には，複数の情報源の間で時間軸方向の変動パターンの類似性を測定する手法を提案している．一方，本論文では，Wikipedia から自動的に抽出された翻訳知識を用いて，日中間の時系列トピックの対応を同定する手法を提案している．また，文献 [5] においては，トピックモデルとして LDA を用い，各日において独立に推定されたトピックを時系列方向に繋げる枠組みを提案している．

6 おわりに

本論文では，日本語および中国語の二言語の時系列ニュースを対象として，トピックモデルのバースト解析の手法を用いて，各日において，バーストするトピックを同定した．そして，時系列に沿って継続的に報道されるトピックに対して，日中間でバーストするトピックの対応をとる手法を提案し，その有効性を示した．

参考文献

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

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

[3] 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*, pp. 532–541, 2012.

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

[5] 芹澤翠, 小林一郎. 潜在トピックの類似度に基づくトピック追跡への取り組み. 第 25 回人工知能学会全国大会論文集, 2011.

[6] 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*, Vol. 7614 of *LNCS*, pp. 239–249. Springer, 2012.

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