日中時系列ニュースにおける トピックの推定と二言語間対応付け*

胡 碩† 高橋 佑介† 牧田 健作† 横本 大輔† 宇津呂 武仁‡ 吉岡 真治§
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‡

1 はじめに

ウェブ上の世界を始めとして，膨大な情報が溢れ，いわゆる情報爆発が起こっている．ウェブ上のニュース記事の上に限っても，同様に多くの情報が流れている．これらのことを背景にして，時系列に沿って，情報を集約したり俯瞰的に把握するための技術が注目されている．例えば，本研究で利用した統計的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技術が確立されてきた．

トピックモデルにおいては，文書が生成される背景において，潜在的に複数のトピックが寄与していることを想定し，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) [2] は，与えられた文書集合から，文書ごとのトピックの確率分布と，トピックごとの語の確率分布を学習する．

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，各日において，DTMによってトピックの分布を推定する．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，日中間でトピックの対応をとる手法を提案する．日中間でトピックの対応をとる際には，Wikipediaの言語間リンクを用いる．日本ニュースと中国ニュースの間では，あるトピックに関する報道期間が異なると，この話題についての関心度が異なる可能性があると言える．例えば，あるトピックについて，日本における報道日数が三日間であるのに対して，中国での報道日数が五日間であるならば，日本においてより，中国においてのほうが関心度は高いと言える．本研究では，このように，日中の時系列ニュースにおける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．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．

2 トピックのモデル化

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) 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) 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．

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，予備実験を通して調整を行い，$\alpha = 0.01$，$K = 10$ とした．

2.2 文書とトピックの対応付け

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．

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

†Shuo Hu, Yusuke Takahashi, Kennsaku Makita, Daisuke Yokomoto 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

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

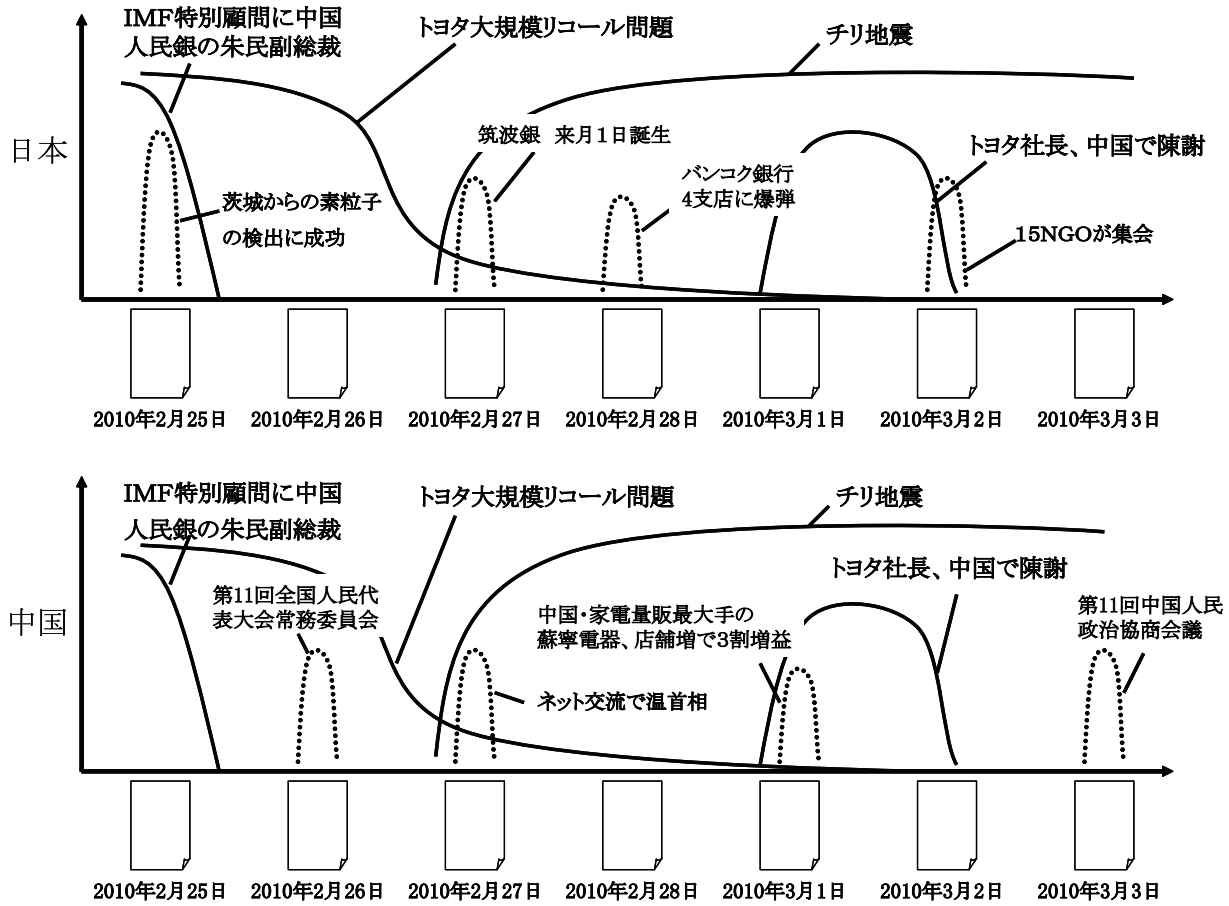

図 1: 日中時系列ニュースにおけるトピックの分析

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．

3 トピックの二言語間対応の推定

3.1 Wikipediaの言語間リンクを用いたエントリの日中対応の抽出

Wikipedia の言語間リンクは，同じタイトルのエントリに対して，Wikipedia の他言語版へのリンクである．Wikipedia の言語間リンクを用いることにより，あるエントリに対して，日本語と中国語の間で，翻訳を行うことができる．そこで，日本語エントリ e_J と中国語エントリ e_C から，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る．ただし，以下では，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，日中対訳語組を抽出する．

例えば，ある日本語エントリを $e_J = \langle J_0, \{J_r^1, \ldots, J_r^l\} \rangle$ とし，中国語エントリを $e_C = \langle T_0, \{S_r^1, \ldots, S_r^k, T_r^{k+1}, \ldots, T_r^h\} \rangle$ とする．ここで，J_0 は日本語エントリのタイトルである．$J_r^1, \ldots, J_r^l$ はこの日本語エントリのリダイレクトである．T_0 は対応している中国語エントリの繁体字タイトルである．$S_r^1, \ldots, S_r^k$ はこの中国語エントリの簡体字リダイレクトである．$T_r^1, \ldots, T_r^h$ はこの中国語エントリの繁体字リダイレクトである．ここで，中国語の簡体字ニュース記事中において，簡体字リダイレクト S_r^i が出現したとすると，e_J と e_C の間の言語間リンクを用いて，S_r^i を含む日中対訳組の集合 $JS(\langle e_J, e_C, S_r^i \rangle)$ を以下のように定義する．

$$JS(\langle e_J, e_C, S_r^i \rangle) = \left\{ \langle J_0, S_r^i \rangle, \langle J_r^1, S_r^i \rangle, \ldots, \langle J_r^l, S_r^i \rangle \right\}$$

3.2 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，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，共有日中対訳語組数を求めて，一定の値 θ_{JC} (本論文では，θ_{JC} を 10 に設定した) 以上の共有日中対訳語組数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，共有され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．

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

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．

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

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

3.3 トピックの日中対応の推定

本節では，まず，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，日中それぞれの言語で DTM によりトピック推定を行い，i 番目の日における日本語トピック集合 TT_J^i および中国語トピック集合 TT_C^i を求める．そして，TT_J^i と TT_C^i の間で，以下の手順により，トピック組の対応を推定する．

まず，日本語トピック $t_J(\in TT_J^i)$ に対して，$P(t_J|d_J) \geq \theta_t$ ($\theta_t = 0.6$) という条件を満たす記事 d_J を集める．同様に，中国語トピック $t_C(\in TT_C^i)$ に対して，$P(t_C|d_C) \geq \theta_t$ ($\theta_t = 0.6$) という条件を満たす記事 d_C を集める．そして，日本語トピック t_J と中国語トピック t_C の間で $P(t_J|d_J) \geq \theta_t$ および $P(t_C|d_C) \geq \theta_t$ を満たす記事組 $\langle d_J, d_C\rangle$ のうち，$DD_{JC}(\theta_{JC})$ 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．

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

表 1: ニュース記事の日中対応の評価結果 (%)

日付	日本語		中国語	
2010 年 2 月 25 日	53.0	(26/49)	54.8	(40/73)
2010 年 2 月 26 日	62.1	(18/29)	62.5	(15/24)
2010 年 2 月 27 日	76.7	(23/30)	88.6	(31/35)
2010 年 2 月 28 日	88.2	(30/34)	87.8	(36/41)
2010 年 3 月 1 日	58.7	(27/46)	54.7	(35/64)
2010 年 3 月 2 日	43.5	(10/23)	40.0	(12/30)
2010 年 3 月 3 日	61.1	(22/36)	25.8	(25/97)
合計	63.2	(156/247)	53.3	(194/364)

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

3.4 時系列トピックの日中対応の推定

時系列トピックの日中対応の推定においては，ある連続期間における日本語と中国語の時系列ニュースにおいて，複数の日に渡るトピックを日中間で対応付ける．連続期間の各日において，DTM モデルで生成した日本語トピック集合の列を $Q_J = TT_J^1, TT_J^2, \ldots, TT_J^n$ とし，中国語トピック集合の列を $Q_C = TT_C^1, TT_C^2, \ldots, TT_J^n$ とする．TT_J^1 は 1 番目の日の日本語トピック集合である．時系列トピックの日中対応の推定は，以下のように行った．

- i 番目の日において，任意の日本語トピック $t_J \in TT_J^i$ に対して，同日において対応付けた中国語トピック $TA_C(t_J, TT_C^i, \theta_t, \theta_{JC})$ を推定する．
- 同様に，任意の中国語トピック $t_C \in TT_C^i$ に対して，同日において対応付けた日本語トピック $TA_J(t_C, TT_J^i, \theta_t, \theta_{JC})$ を推定する．

以上の手順を，$i = 1, \ldots, n$ について行うことにより，連続期間の全体を通して時系列トピックの日中対応を推定する[2]．

[2]実際には，DTM においては，連続期間中の各日の間でのトピックの対応の情報を利用することが可能であるが，本論文の評価実験においては，各日の間のトピックの対応の情報は利用せず，各日別に日中間のトピック対応の推定を行った．ただし，評価実験において推定された日中間のトピックの対応関係のうち，各日の間のトピックの対応と矛盾するものはなかった．

表 2: 各日におけるトピックの日中対応の出力およびトピック単位での評価結果

分類	話題	2010 年 2 月				2010 年 3 月		
		25 日	26 日	27 日	28 日	1 日	2 日	3 日
日中共通 (正しい トピック対応)	トヨタリコール 問題	有	有	有				
	トヨタが 中国で謝罪					有		
	チリ地震			有	有	有	有	有
	IMF 特別顧問に 中国人民銀の副総裁	有						
日中共通 (誤りの トピック対応)	自国経済について のニュース	有	有	有	有	有	有	有
中国語側 からのみ 出力し，誤りの トピック対応	温首相が ネットで交流			有 (中国語側のみ 適切なトピックを推定. 中国語 10 記事が 日本語 1 記事に対応)				
トピック単位での評価結果: 日本語トピック 80.0%(4/5) 中国語トピック 66.7%(4/6)								

4 分析および評価

ニュース記事の日中対応推定結果の評価は，2010 年 2 月 25 日から，3 月 3 日までの一週間で行った．この一週間での日本語ニュース記事数は 3,278 件，中国語ニュース記事数は 6,110 件であった．一週間におけるニュース記事の日中対応の評価結果を表 1 に示す．この結果において，大きく精度を下げている要因は，日中両国における自国経済についてのニュースである．これらのニュースは，日中両国間において共通の内容とはならないにも関わらず，数値や経済用語等が共通に出現することが要因となって，日中間において誤った記事対応として多数出力されている．実際に，これらの自国経済についてのニュース記事を除外した後算出した日中間記事対応の精度は，90%以上となっていた．これらの自国経済ニュースは，年間を通じて定常的に報道されるため，今後は，トピックのバースト [3] を検出する手法 [5] を併用することにより，除去することが適切であると考えられる．

次に，表 1 において評価対象となった日本語ニュース記事 247 記事，中国語ニュース記事 364 記事を対象として DTM モデルを適用し生成したトピックに対して，日本語と中国語の間でトピックの対応の推定を行った．その結果の抜粋を表 2 に示す．この結果から分かるように，全ての日において，自国経済についてのニュースに相当するトピックが日中間で対応付けられており，トピック単位ではこれらはすべて誤りの対応となった．その他，中国でのみ盛り上がったトピック「温首相がネットで交流」についても，中国語側からのみトピック対応が出力されたが，誤りのトピック対応と判定した．その他の 4 種類のトピック対応については，適切なトピック対応を出力した．

5 関連研究

文献 [4] 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．一方，時系列方向のトピックのつながりを理論的にモデル化したトピックモデルとして，On-line LDA [1] などがある．

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，各日において，DTM によってトピックの分布を推定した．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] L. ALSumait, D. Bardara, and C. Domeniconi. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proc. 8th ICDM*, pp. 3–12, 2008.

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

[4] 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, pp. 111–114, June 2011.

[5] 高橋佑介, 横本大輔, 宇津呂武仁, 吉岡真治. ニュースにおけるトピックのバースト特性の分析. 情報処理学会研究報告, 第 2011 巻, 2011.