Web ビッグデータからの地域研究情報抽出の試み（第二報）

原正一郎（京都大学東南アジア地域研究研究所）・山田太造（東京大学史料編纂所）
石川正敏（東京成徳大学経営学部）・白井圭佑（京都大学情報学研究科）
亀田尭宙（国立歴史民俗博物館研究部）・森信介（京都大学学術情報メディアセンター）

近年の地域研究はディシプリン重視の思潮が強く，特定地域の特定研究分野を個別的に捉えようとする傾向にあるが，本来は研究分野や個別地域の枠を超えて，地域の全体像を読み解くことを目的としている．地域情報学は，地域研究の本来の目的に対応する研究分野で，断片的データを情報学的に組み合わせて地域の全体像を再構築するボトムアップアプローチを旨とする．しかし，データの欠落や，研究手法の相違によるデータの種類・質・粒度等の不整合により，データを組織化して地域の全体像を再構築することは出来なかった．そこで，本研究はトップダウン的な情報学手法の確立を目指す．ここでは，Web 上のビッグデータから研究のヒントとなる情報を自動的に抽出・整理・可視化する情報システムの開発を試みる．本稿は，その第二報である．

A Trial to Extract Area Study Information from Web Big Data (Second Report)

Shoichiro HARA (Center for Southeast Asian Studies, Kyoto University)
Taizo YAMADA (Historiographical Institute, University of Tokyo)
Masatoshi ISHIKAWA (Faculty of Business Administration, Tokyo Seitoku University)
Keisuke SHIRAI (Graduate School of Informatics, Kyoto University)
Akihiro KAMEDA (Research Department, National Museum of Japanese History)
Shinsuke MORI (Academic Center for Computing and Media Studies, Kyoto University)

Recent area studies tend to focus on particular research domains, that is, try to understand a particular area from the perspective of a particular research domain. However, original area studies try thoroughly to grasp an area beyond boundaries of countries and research domains. Area informatics is a scientific field corresponding to this original area studies. It aims to reconstruct an overall picture of an area by a bottom-up approach to assemble fragmented data. However, due to lack of data and inconsistencies in data type, quality, and granularity depending on differences in research methods, it was not possible to organize the fragmented data and reconstruct an overview of an area. Therefore, this study is to establish a top-down approach for area studies, that is, tries to develop an information system that automatically extracts, organizes, and visualizes big data on the Web and to create hints for researches. This paper is its second report.

１．まえがき

近年の地域研究はディシプリン重視の思潮が強く，特定地域の特定研究分野を個別的に捉えようとする傾向にある．しかし本来の地域研究は，研究分野や個別地域の枠を超えて，例えば「中国とは何か」というように，ある地域に関する全体像を再構築する（ある地域研究者は「読み解く」と表現する）ことを目的としている．地域情報学は，地域研究の本来の目的に対応する情報学研究分野であり，断片的データから地域の全体像を組み上げるボトムアップアプローチを旨とする．地域情報学は，地域研究資料のデータベース化を主要な成果とし，さらにデータベース同士の横断的連携も実現するなどして，これまで地域研究の進展に貢献してきた[1].

しかし，ピースの不足やピース同士の形が合わなければジグソーパズルが完成しないのと同様に，「地域×研究分野」の偏りが著しい地域研究では，データの不足や，研究手法の相違によるデータの種類・質・粒度等による不整合性が多く，断片的なデータから地域の全体像を再構築できなかった．したがって，「地域×研究分野」というピースを組み上げる旧来型の方法とは異なる，新しいパラダイムの構築が必要である．

そこで本研究では大量・多様・リアルタイム性の高いビッグデータを利用したトップダウン的なアプローチを試みる．ここでは，「ビッグデータには地域の全体像を読み解くに充分な情報（ピース）が含まれている」という前提に立ち，そこから地域研究に有用な情報を抽出し（ピース選択），それらを組織化して（パズルの完成），新しい学知の発見に繋げる（読み解く），という一連の手法を開発するアプローチである．

ビッグデータは，地域研究の主要なテーマである紛争・災害・政変・選挙等の発生時において流通量が顕著に多くなる．しかし，ビッグデータが包含する関連データの量や多様性は，一人の研究

者が手作業で一生をかけて収集できるそれを遥かに凌駕している．地域研究研究者の多くが協力してビッグデータを使いこなすことができるようになれば，地域研究をより深化させ，その可能性を飛躍的に拡大させることができよう．

そこで本研究では，ビッグデータから地域研究のヒントとなる情報を自動的に抽出して可視化する情報システムの開発を試みる．これにより，情報学を主体とした地域研究というインターネット時代に対応した新しい研究の方向性を模索する．

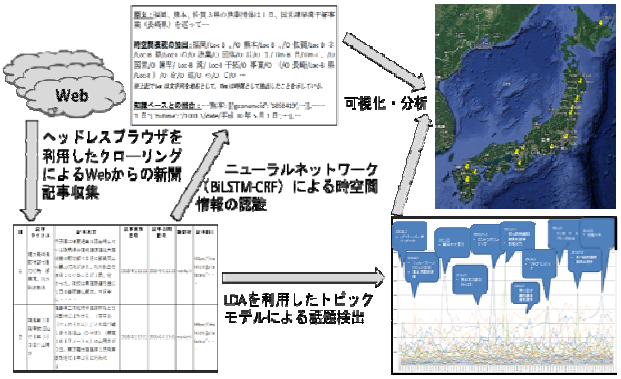

図 1. システムの概要
Figure 1 Overview of the developed system

2．方法

Web から収集した新聞記事（以下，記事）を分析する事例は地域研究においても散見される．あるいは，災害時における最適な避難計画を策定するために，Social Media などのビッグデータを活用して，住民の行動様式を分析する社会学研究の事例などもある．これらの研究では，対象とする情報源を限定した上で，キーワードを巧みに利用して適切な記事を収集する．分析対象が定まっているので，この手法は有効である．

しかし，地域の特徴・動向を総体的に把握する場合，キーワードを事前に設定することはできない．そこで本研究では，Web 上の新聞記事（以下，記事）を対象として，①Web scraping によるデータ収集，②機械学習に基づく形態素解析と時空間表現の認識，③Topic Model による話題抽出，④可視化という処理の組み合わせを試みた（図 1）．記事を対象と利用した理由は，データ収集と分析の容易さを優先させたためである[2].

2.1Web からの新聞記事収集

新聞社の自社サイト上の記事を収集対象とした．ただし，新聞社の多くが記事の有料化を進めているため，本研究では無料で閲覧可能な部分だけを収集対象とした．Web scraping 技術を用いて毎日新聞・朝日新聞・読売新聞・AFP（英語版）から記事を収集するとともに，各記事のタイトル・本文・更新日時・公開日時・URI を抽出した．これらを記事情報として CSV 形式のテキストファイル（UTF-8，BOM なし）に編集し，以降の処理で使用した．

2.2 新聞記事からの時空間情報認識

記事中の時空間表現を認識するため，最初に形態素解析器 KyTea [3]により，記事中の文を単語列に分割した．KyTea は，現代日本語書き言葉均衡コーパス（BCCWJ）[4] のコアデータ（557,281 文，単語分割・品詞付与済み）により学習しており，同じ分野のテストデータに対する精度は 98%以上（3,024 文のテストデータ）であった．

次にニューラルネットワークにより，各単語を空間表現（図 2 ではタグ Loc で指示），絶対的な時間表現（同，TimA），相対的な時間表現（同，TimR），それ以外の単語（同，O）に分類する．同時に，時空間表現に関する単語については，開始（同，B）または継続（同，I）であるかを識別する．採用したニューラルネットワークは，双方向長期短期記憶ネットワーク（Bi-directional Long Short Term Memory; BiLSTM）と条件付き確率場（Conditional Random Fields; CRF）を組み合わせた BiLSTM-CRF[5]である．BiLSTM-CRF の概要を図 2 に示す．ここで，矩形は BiLSTM を，台形は CRF 層を，それぞれ表している．BiLSTM-CRF は単語列を入力として受け取り，対応するタグ（Loc-B, Loc-I, TimA-B, TimA-I, TimR-B, TimR-I, O）のいずれか 1 つを推定する．

時空間表現に分類された単語は，辞書を参照して，絶対的な時間表現は西暦年月日に，空間表現は緯度・経度に変換した．

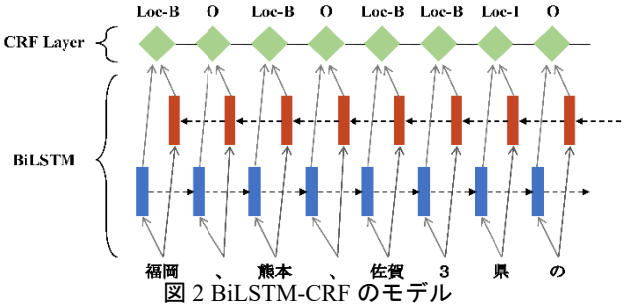

図 2 BiLSTM-CRF のモデル
Figure 2 Schema of BiLSTM-CRF model

2.3 新聞記事からの話題抽出

収集した記事から話題を自動的に検出するために，Topic Model の 1 つである Latent Dirichlet Allocation (LDA) [6]を用いた．LDA では 1 つの文書に複数のトピックが存在すると仮定しており，文書全体のトピック分布および記事ごとのトピックの分布をモデル化する．図 3 は本研究で用いた LDA のグラフィカルモデルを示す．ここで，青塗りの円は観測変数，白塗りの円は未知変数を示す．矩形は繰り返しを示し，右下の数字は繰り返しの回数を示す．ここで w は唯一の観測変数である抽出された用語を示す．z はトピック，θ は記事ごとのトピック分布，ϕ はトピックごとの用語分布を示す．α および η はそれぞれ θ および ϕ

のパラメータであり，LDA としてはパラメータのパラメータであることからハイパーパラメータと呼ばれる．

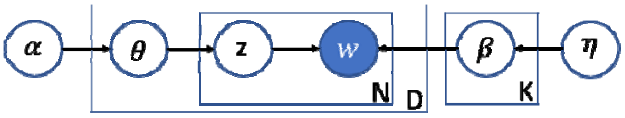

図 3 LDA のグラフィカルモデル
Figure 4 Graphical model for LDA

3．結果と考察

3.1 時空間情報認識

BiLSTM-CRF は単語分割済みのデータを入力とするが、推定の際には単語内の文字情報も利用する。 BiLSTM-CRF のパラメータは，上記の時空間表現に関するタグを人手で各単語に付与したデータ 3,488 文（学習データ 2,789 文，開発データ 349 文）から推定し，テストデータ 350 文に対して精度を測定した．その際の学習曲線を図 4 に示す．この推定実験において，BiLSTM-CRF の認識精度は，学習データとして 2,789 文全てを用いた場合，次空間表現の認識精度（F 値）で約 83.8%であった． この 83.8%という認識精度は BiLSTM-CRF の有効性を示すものであるが，実用という観点からは十分であると断言できない．図 4 の学習曲線は，学習データの追加により，更なる精度向上が期待出来ることを示している．

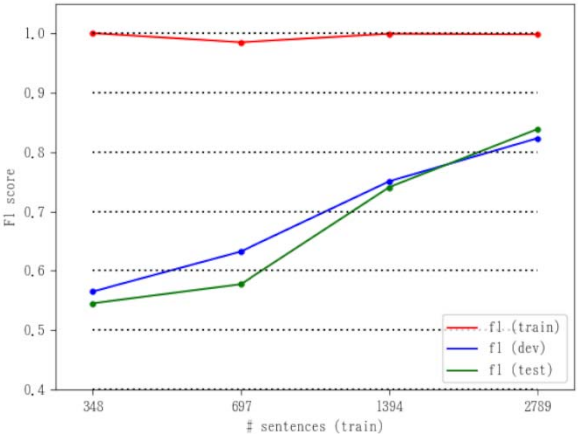

図 4 BiLSTM-CRF の学習曲線.
Figure 4 Learning curves of BiLSTM-CRF model

3.2 地名表現の曖昧性解消

BiLSTM-CRF によって抽出された空間表現には，同名異地名などの曖昧性が存在する[2]．この問題を解消する試みとして，同一記事中の地名は相互に近傍に存在するという前提にたち，各地名間の総距離が最小になるように地名を選択した．具体的には，記事中の全ての地名が指す点を丁度一度ずつ巡る経路のうち，総距離が最小となる地名の組合せを選択する，つまり巡回セールスマン問題として曖昧性の解消を図った．

記事中の地名の数を n とした時，総当たり法による時間計算量は O(n!)となり，n が大きくなると現実的ではない．そこで，局所探索アルゴリズムの一種である 2-opt[7]の利用を試みた．これは，現在の経路から 2 地点を選択し，それらを繋ぎ変えた際に総移動距離を短縮可能であるならば，その部分経路を採用する操作を繰り返す手法である．実装では，記事中に存在する全ての 2 地点について繋ぎ変えを考慮するため，時間計算量は O(n^2)となる．図 5 に n を変化させた際の両手法の計算時間を示す．総当り法では 9 以上の n に対しては計算時間が非常に大きくなっているが，2-opt では顕著な計算時間の増大を見せないことがわかる．

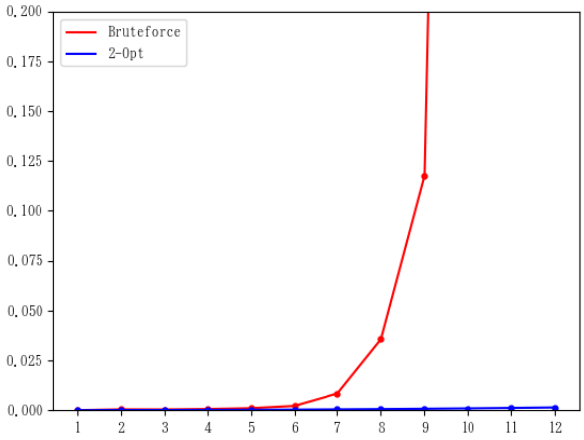

図 5 総当り法と 2-Opt の計算時間.
Figure 5 Computation time of brute force and 2-opt

3.3 話題抽出

LDA の学習データとして 2010 年から 2015 年までの 6 年間の毎日新聞記事（CD-毎日新聞 2010~2015 年データ集, 606,924 記事）を使用した．さらに，学習後の LDA モデルを用いて，2018 年 12 月から 2019 年 10 月までに収集した新聞記事（31,417 記事）を用いて再学習を行った．一つ以上の連続する名詞およびその直後に接尾辞が続くものを用語として抽出した．学習では，LDA におけるトピック数を 200， α および η をそれぞれ 0.1 と設定し，パラメータの推定には変分ベイズ法を用いた．

図 6 に LDA により抽出したトピックのうち，上位 20 トピックを示す．そのうち上位 5 件のトピックに含まれる用語は次の通りであった．

- トピック 193：こと，ため，問題，政府，…
- トピック 94：こと，人，もの，中，今，…
- トピック 109：発表，ため，影響，こと，被害，…
- トピック 56：選手，優勝，決勝，チーム，出帳，…
- トピック 173：逮捕，事件，男，県警，疑い，…

他には,震災に関するトピック（トピック 21）や教育に関するトピック（トピック 17）,ロシアに関するトピック（トピック 60）等があった.このうち,震災に関するトピックに含まれる用語を

Word Cloud を用いて可視化したところ,図 7 に示すような結果が得られた.

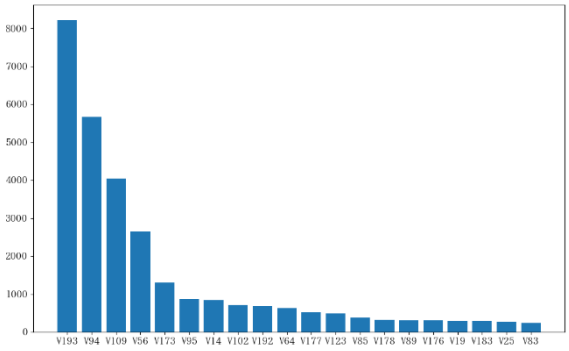

図 6 LDA によるトピック検出.
Figure 6 Topic detection by LDA

今回の LDA では，Web から収集した記事の解析に際して，KyTea により推定した品詞のみを基に用語抽出を行っている．KyTea により推定した品詞に加えて，BiLSTM-CRF により抽出した時空間表現（Loc-B, Loc-I, TimA-B, TimA-I, TimR-B, TimR-I, O）も用語の対象とすることで，時空間表現に関するトピックをより明確に抽出できる可能性がある．

そこで，収集した記事を対象に，KyTea のみの場合と，BiLSTM-CRF を組み合わせた場合を比較したところ，抽出されたトピックとそれに含まれる用語に顕著な違いが見られなかった．収集した記事ではデータ量が不足しており，BiLSTM-CRF による時空間表現の抽出量が十分ではなかったことが原因の一つとして考えられる．そこで，より多くの記事（例えば 6 年分の CD データ）に BiLSTM-CRF を適用したら，時空間表現に関するトピックの検出が改善されるかを，実験により検証することが課題となる．

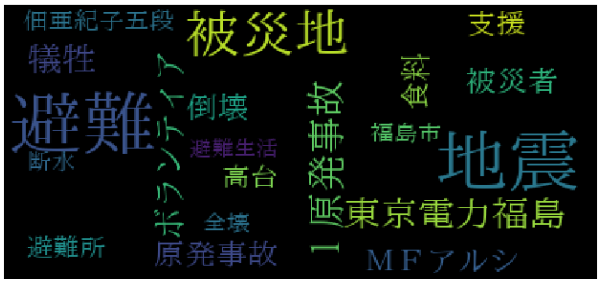

図 7 Word Could による可視化.
Figure 7 Visualization by Word Cloud

3.4 トピックごとのイベント検出

LDA においては，複数のトピックから生成された語の集まりが一つの文書を構成すると考える．つまり，文書を構成するトピックの割合を比較することで，文書の類似性を推定できる．そこで，特定のトピックを多く含む記事を時系列的に追うことで，何時・どのようなイベントが社会の関心事になっていたのかを抽出する試みを行った．アルゴリズムは以下のとおりである．

① 着目したいトピックを決める（以降，トピック k とする）．
② トピック k の割合が最大となっている記事を 1 日につき 1 記事選び出す．
③ 調査期間 t_{all} 中において，選択した記事のトピック k の割合を平均し，

（1+全期間中のトピック k の割合平均）/2

を閾値と定め，この閾値を越える記事が 3 日以上連続している期間集合 $T=\{t_1, t_2, \ldots t_i, \ldots\}$ を検出する．なお，上式において，1 が割合の最大値であるので，これは，最大値とトピック割合の平均値の間にそのトピックが強いとみなすための閾値を置いていることになる．また，大きなイベントであれば，3 日以上は話題になっているだろうという仮定を置いた．

④抽出した各期間中の記事に共通して出現している（複合）名詞を抽出する．その際，各記事の先頭 30 名詞のみに着目する．これは，単文書要約で強力なベースラインとして用いられるリード法[8]を参考に，重要な用語は記事の冒頭に現れるという仮定によっている．その名詞集合の和集合を調査期間全体で集計したものをイベントのラベルの候補とする（$N=\{n_1, n_2, \ldots, n_j, \ldots\}$）．

⑤ 調査期間中のある名詞の出現日数を $d(n_j, t_{all})$ とし，逆数 $1/d(n_j, t_{all})$ を一日あたりのその名詞のスコアとする．そして，連続期間 t_i ごとに，当該の名詞 n_j が含まれている

日数×$1/d(n_j, t_{all}) = d(n_j, t_i)/d(n_j, t_{all})$

を計算し，それを最大にするような名詞 n_j を連続期間 t_i のラベルとする．これは文書内での各語の重要語を評価する tf-idf の発想に基づいており，トピック内の記事に共通して出現する語彙を軽く，期間限定で出現する語彙を重くすることを意図している．

これにより，各トピックの観点ごとに，そのトピックにとって大きなイベントがあった期間とそのラベルの抽出を試みる．2018 年 6 月 1 日～2018 年 11 月 2 日の記事に適用したところ，台風や大相撲といったトピックで特に分かりやすい結果を得た．以下に概要をまとめる．

台風に関わる例を図 8 に示す．なお，オレンジの点は抽出されたイベント期間であることを示す（図 9 も同様）．ここでは，「台風 5 号」「台風 6 号」「台風 7 号」「台風 8 号」のような台風の名前がラベルとして抽出できた．ただし，「台風 19

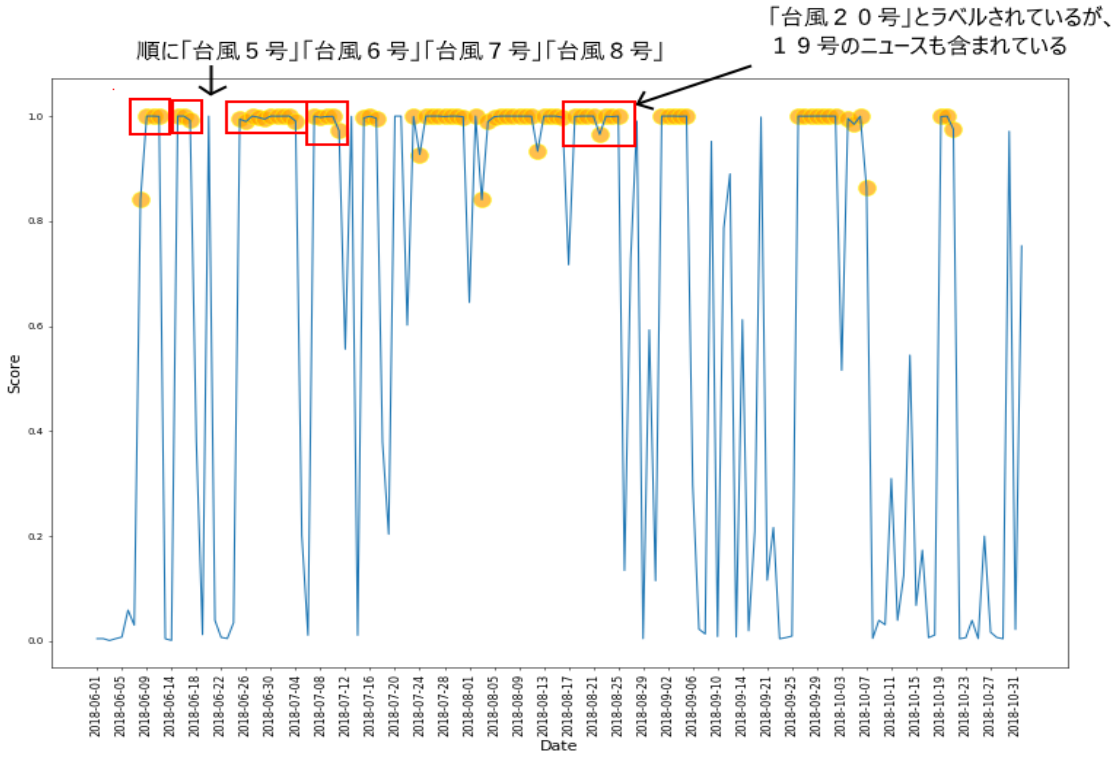

図 8 トピックに着目したイベントの抽出（台風に関わるトピック）
Figure 8 Event detection based on typhoon-related topic of LDA

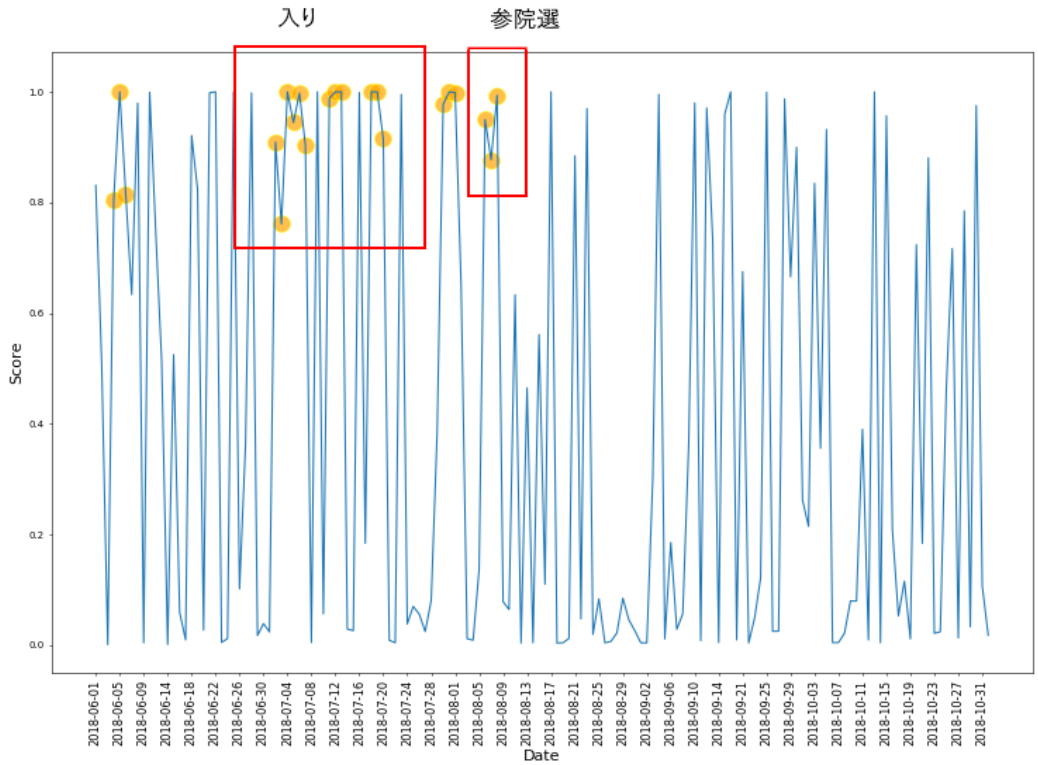

図 9 トピックに着目したイベントの抽出（国内政治に関わるトピック）.
Figure 9 Event detection based on domestic-politics-related topic of LDA

号」と「台風 20 号」については，両方とも「台風 20 号」とラベル付けされてしまった．これは，2 つの台風が連続して日本に接近したため，記事として取り上げられる期間が重なり，一つの期間のイベントとしてラベル付けされためである．

大相撲に関わるトピックについても同様の結果が得られ，「大相撲名古屋場所」，「大相撲秋場所」，「大相撲九州場所」といったラベルを抽出できた．

国内政治に関わるトピック（図 9）では，公職選挙法改正に関わる審議が参議院や衆院憲法審査会で開始されるという記事が抽出された．しかし，「審議」というサ変動詞の語幹を複合名詞の要素に含めていなかったため，複合名詞の取得に失敗し「入り」とラベル付けされていた．なお，参院選を 1 年後の 2019 年に控えて，各党がそれに向けて動き始めるといった細かい動向も，2018-08-06 から 2018-08-08 の記事において，「参院選」というラベルで抽出できた．しかし，前述の審議入りの記事と期間が重複したために，「参院選」としてラベル付けされるはずのイベント期間が短くなってしまった．

以上の結果をまとめると，台風やスポーツシーズンのような比較的短期間の連続的なイベントについては，期間とラベルを効率的に抽出できた．しかし，イベントの継続期間が重複する場合には，一つの大きなイベントとして抽出してしまうという問題点等も明らかになった．

トピックの割合に大きな偏りがあることも問題である．実際，今回のアルゴリズムで抽出できたイベントに関連するトピックは，全トピック数 200 中の 32 トピック（全体の 16%）にとどまった．他方，「こと」あるいは「問題」といった一般的な語が多く含まれるトピックも多いため，役にたたないイベントも抽出されている．

これら以外にも，裁判のように数カ月や数年間におよぶトピックの場合，出現間隔が疎となるため，③で仮定した「連続した 3 日間」を満たさず，イベントとして正しく抽出できなかった．長期間に渡って散発的に出現するトピックについても抽出できるアルゴリズムを考える必要がある．

４．謝辞あとがき

本研究は，JSPS 科研費 JP16H01897 の助成，京都大学研究連携基盤未踏科学研究ユニット経費，京都大学東南アジア地域研究研究所グローカル情報ネットワーク経費を受けたものである．

参考文献

[1] Shoichiro Hara: Area Informatics – Concept and Status –, In: Ishida T. (eds) Culture and Computing. Lecture Notes in Computer Science, vol 6259. Springer (2010), DOI https://doi.org/10.1007/978-3-642-17184-0_17

[2] 原正一郎，山田太造，石川正敏，白井圭佑，亀田尭宙，森信介：Web ビッグデータからの地域研究情報抽出の試み，じんもんこん 2018 論文集，pp. 365-372，2018.

[3] Neubig, Graham, Yosuke Nakata, and Shinsuke Mori: Pointwise prediction for robust, adaptable Japanese morphological analysis, Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011.

[4] Maekawa, Kikuo, Makoto Yamazaki, ToshinobuOgiso, Takehiko Maruyama, Hideki Ogura, WakakoKashino, HanaeKoiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den: Balanced corpus of contemporary written Japanese, Language resources and evaluation 48.2 (2014): 345-371.

[5] Huang, Zhiheng, Wei Xu, and Kai Yu: Bidirectional LSTM-CRF Models for Sequence Tagging, arXiv preprint arXiv:1508.01991 (2015).

[5] 前川喜久雄: 代表性を有する大規模日本語書き言葉コーパスの構築，人工知能学会誌, vol. 24, No. 5, pp. 616-622，2009, http://pj.ninjal.ac.jp/corpus_center/bccwj/

[6] D.M.Blei, A.Y.Ng, and M.I.Jordan: Latent Dirichlet Al- location, Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.

[7] Croes, Georges A. A method for solving traveling-salesman problems. Operations research 6, no. 6 (1958): 791-812.

[8] 高村大也: 文書要約のための数理的手法．経営の科学 62(11), pp.711-716, 2017.