

Web ビッグデータからの地域研究情報抽出の試み

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 堯宙（京都大学 東南アジア地域研究研究所）・森 信介（京都大学 学術情報メディアセンター）

インターネットの普及に伴い、地域研究に関わる大量の情報が Web 上で流通するようになった。文字メディアに限定しても Web 上のデータ量は膨大である。研究に必要な情報を網羅的に抽出することは困難となりつつあり、抽出した全データをヒトが読んだり分析したりすることはもはや不可能である。本稿では、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して可視化する情報システムの開発を通じ、インターネット時代に適応した情報学による地域研究の方向性を模索する。

A Trial to Extract Area Study Information from Web Big Data

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 (Center for Southeast Asian Studies, Kyoto University)

Shinsuke MORI (Academic Center for Computing and Media Studies, Kyoto University)

With the spread of Internet, a great deal of information related to area studies has been distributed on the Web. Even if the information is limited to text media, the amount of data on the Web is still enormous. It is becoming more difficult to extract the information necessary for research, and is unattainable to read and analyze all the extracted data as before. This paper will explore the potential of new research directions for area studies by information science which is adapted to the Internet age. This will be done through the development of an information system that automatically extracts and analyzes big data on the Web. This system will also deliver visual products from the analyses that can serve as hints for the future development of research in area studies.

1. まえがき

地域研究は人文科学・経済学・自然科学・工学・保健医学等を包含する学際研究領域である[1]。その中でも主要な研究領域である人文科学に注目すると、これまでの主要な研究資源は史資料・研究論文・新聞・書籍・雑誌等の文字メディアであり、地域研究者は文字を読み解きながら研究を進めていた。ところがインターネットの普及に伴い、地域研究に関わる多種・大量の情報が Web 上で流通するようになった。この傾向は紛争・災害・政変・選挙等の発生時において顕著である。文字メディアに限定してさえ Web 上のデータ量は膨大であり、研究に必要な情報を網羅的に抽出することすら困難となりつつあり、まして抽出した全データをヒトが読んだり分析したりすることは、もはや不可能である。

そこで本研究では、Web 上の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。

2. 方法と結果

Web ビッグデータから有用な地域研究情報を抽出する手法開発の第一段階として、本研究では新聞記事を対象とした。その理由は、①Twitter や Facebook 等の Social Media Site に比べると、新聞記事のテキストは文法的にも語彙的にも整っているため機械的な分析が容易であること、②新聞は政治・経済・文化等の情報を網羅的に得るうえで有効な情報源であること、③新聞社の多くは、記事を紙媒体だけではなくデジタルデータとし

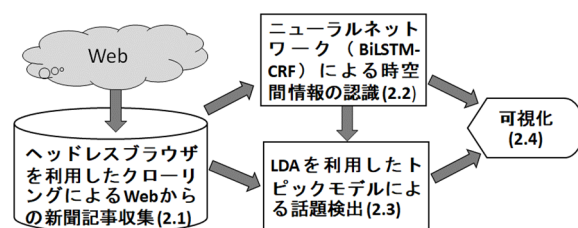


図1 開発システムの概要

Figure 1 Schema of the developed system

て Web で公開しているため、データ収集と処理が容易なためである。

本研究では、図 1 に示す手法を組み合わせることにより、Web からの新聞記事収集（2.1 節で説明）に始まり、新聞記事中の時空間情報に関わる語彙の認識（2.2 節）、新聞記事内容からの主題抽出（2.3 節）、さらに可視化（2.4 節）までを実現している。

2.1 Web からの新聞記事収集

新聞記事を Web で公開する場所としては、新聞社の自社サイトに加えて、ポータルサイトやキュレーションソフトなどが挙げられる。複数社の新聞記事を収集するならばポータルサイトが効率的であると考えられる。しかし、ポータルサイトで公開される新聞記事は、新聞社サイトで公開されている記事の一部だけあったり、新聞社サイトの記事を Web ページに転送しているだけであったりする。そこで、本研究の新聞記事収集システム（以下、本新聞記事収集システム）では、新聞社の自社サイト上の新聞記事を収集対象とした。ただし、新聞社の多くが新聞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。今回の収集対象は、毎日新聞、朝日新聞、読売新聞、AFP（英語版）である。本新聞記事収集システムが新聞社サイトから収集する情報は以下の通りである。

- 記事タイトル
- 記事本文
- 更新日時
- 公開日時（ただし、公開日時の記載のない新聞記事は更新日時と同一とした。以下②参照）
- 記事 URL（新聞記事本文が掲載されている Web ページの URL）

以上に加えて、本新聞記事収集システムでは、データ管理用の ID（単純な通し番号）と、新聞記事提供元を識別するために新聞社名を記録した。新聞記事の収集手順は以下の通りである。

- ① 新聞記事の更新リストを取得する。更新リストは、毎日新聞と朝日新聞では RSS を利用し、RSS を提供していない読売新聞と AFP では新着記事の一覧を記載している Web ページから、該当する情報を取得した。更新リストの各要素は、記事タイトル、更新日時、記事 URL の組である。図 2 に取得した記事の例を示す。
- ② 手順①で取得した更新リストの要素にある記事タイトルと更新日時から、対象要素がすでに収集された新聞記事であるかを確認する。その結果、これまでに収集されていない新規の新聞記事である場合は手順③を実行し、そうでなければ更新リストの次の要素を処理する。

242. 諫干: 3 県漁業団体、基金案受け入れ 開門せず協議継続。福岡、熊本、佐賀 3 県の漁業団体は 1 日、国営諫早湾干拓事業（長崎県）を巡って堤防開門を強制しないよう国が漁業者に求めた訴訟で、開門せずに漁業振興基金を設ける和解案の協議を継続するよう（中略）。佐賀県有明海漁協の徳永重昭組合長は「3 県が歩調を合わせることで有明海再生に弾みがつくのではないか、和解協議は何とか続けてもらいたい」と話した。【池田美欧】
2018-05-01 21:38:57 2018/5/1 21:38:57 mainichi
<https://mainichi.jp/articles/20180501/k00/00e/020/345000c>

図 2 収集した新聞記事の例

Figure 2 An example of aggregated article

- ③ 手順②で新規の新聞記事と判断された場合、更新リストの要素にある記事 URL を参照して、新聞社のサイトから新聞記事の本文が記載されている Web ページをダウンロードする。次に、ダウンロードした Web ページから新聞記事本文に該当する HTML 要素の内容（記事本文）を取得する。Web 上で公開される新聞記事には、記事本文冒頭に付けられる記事タイトルの他に、記事本文中に小見出しが追加されていることがある。しかし、紙場媒体の記事の多くは、そのような小見出しが付くことはない。したがって、今回の新聞記事収集では、記事本文中の小見出しは収集しなかった。また、毎日新聞と読売新聞では、各新聞記事の本文が記載されている Web ページの meta 要素に新聞記事の公開日時が記載されているため、記事本文の収集に合わせて取得する。最後に、手順②と③で収集したデータを新規の新聞記事としてデータベースに登録する。更新リストの要素がなくなるまで、手順②から③を繰り返す。
- ④ 記事収集対象の新聞社のサイトへのアクセスの負荷を考慮して、適当な時間をあけて手順①から③を繰り返し、連続した新聞記事収集を実現する。

本新聞記事収集システムは、収集した新聞記事を新聞社ごとに CSV 形式のテキストファイルに

ID	記事 タイトル	記事本文	記事変更日時	記事公開日時	新聞社	記事URL
242	諫干: 3 県漁業団体、基金案受け入れ 開門せず協議継続	福岡、熊本、佐賀 3 県の漁業団体は 1 日、国営諫早湾干拓事業（長崎県）を巡って堤防開門を強制しないよう国が漁業者に求めた訴訟で、開門せずに漁業振興基金を設ける和解案の協議を継続するよう求める共同文書を公表した。これまで開門を求めてきた・・・	2018/5/1 21:38	2018/5/1 21:38	mainichi	https://mainichi.jp/articles/20180501/k00/00e/020/345000c
243	三井住友と大和証券の系列の資産運用会社、統合で調整	三井住友フィナンシャルグループ（F G）と大和証券グループは、系列の資産運用会社を来春をめどに統合する方向で調整に入った。規模拡大でシステム投資などの効率化を進め、運用力の強化を目指す。統合を検討しているのは、三井住友 F G が 6 0 % 出資する三井住友アセットマネジメント・・・	2018/5/1 21:36	2018/5/1 21:36	mainichi	https://mainichi.jp/articles/20180501/k00/00e/020/345001c

図 3 収集した新聞記事の CSV 出力例

Figure 3 Output examples of aggregated articles in CSV format

変換して次の処理 (2.3) に渡している。このテキストファイルの文字コードは、utf-8 (BOM なし) である。出力例を図 3 に示す。

本新聞記事収集システムは Python 3.6 で記述し、主なライブラリとして、RSS の構文解析用に feedparser, HTML 文書の解析用に BeautifulSoup4 を使用した。収集した新聞記事の管理用データベースには SQLite3[2]を使用している。近年、Web サイトから Web ページを取得するときに cookie や JavaScript などの前処理をしなければ、必要な情報が含まれた Web ページを得ることができないケースが増えている。新聞社の自社サイトも同様であり、このような問題を解決する方法として、本新聞記事収集システムにおいてもクロールで広く使われているヘッドレスブラウザを用いて、Web サイトから得る Web ページのレンダリング結果を取得して新聞記事本文を取得している。本新聞記事収集システムでは、ヘッドレスブラウザに PhantomJS[3]を使用している。

本新聞記事収集システムで収集した新聞記事は、毎日新聞 24,851 件 (2017 年 12 月から 2018 年 9 月まで)、朝日新聞 13,903 件 (2018 年 6 月から 2018 年 9 月まで)、読売新聞 10,552 件 (2018 年 6 月から 2018 年 9 月まで)、AFP12,137 件 (2018 年 6 月から 2018 年 9 月まで) である。機器の故障や新聞社サイトの仕様の変更などにより未収集の期間もあるがおおむね上記の期間連続して新聞記事を収集できた。

2.2 新聞記事からの時空間情報認識

前節の手順により Web から収集した新聞記事中の時空間表現を認識するために、機械学習に基づく手法を開発した。具体的には、まず新聞記事を構成する文を単語に分割し、次にニューラルネットワークを利用して、各単語を時間表現に関する単語 (図 4 ではタグ Tim で指示)、空間表現に関する単語 (同, Loc)、時空間表現以外の単語 (同, O) に識別すると同時に時空間表現に関する単語については開始 (同, B) あるいは継続 (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、双方向の長期短期記憶ネットワーク (Bi-directional Long Short Term Memory; BiLSTM) [4]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 は固有表現認識のタスクにおいて高い認識精度を実現することで知られている。また、BiLSTM-CRF の処理単位として、文字が使用されることもあるが、本研究では入力文は形態素解析器 KyTea [6]によって単語分割されていると想定し、処理単位としては単語を使用する。

BiLSTM-CRF モデルの全体図を図 4 に示す。ここで矩形は BiLSTM (下側は文頭から文末に単語列を読み込む LSTM であり、上側は逆向きに読

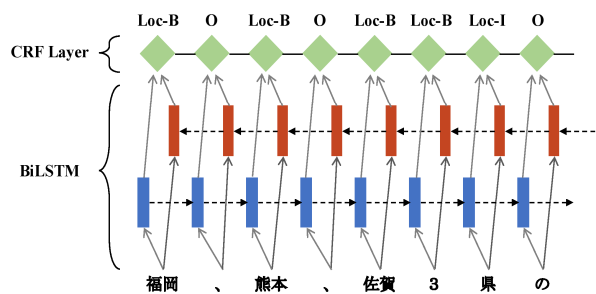


図 4 BiLSTM-CRF のモデル。
Figure 4 Schema of BiLSTM-CRF model

み込む LSTM である) を、緑の台形は 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 BIO タグ (Tim-B, Tim-I, Loc-B, Loc-I, O) の 1 つを推定する。

形態素解析器 KyTea は、現代日本語書き言葉均衡コーパス (BCCWJ) [7] 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。

時空間表現認識器 BiLSTM-CRF のパラメータは、人手で各単語に BIO タグを付与した BCCWJ の白書 (OW) の学習データから推定した。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 5,000 文を用いた場合、次空間表現の認識精度 (F 値) で約 90% (398 文のテストデータ) であった。この 90% 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。また、学習データである BCCWJ の白書 (OW) と適用先である新聞記事のドメインの差異による精度の変化も想定される。

学習データの追加による精度向上の可能性を調べる為に、学習データ 5,000 文に対して、そのサイズを 1/64 (78 文), 1/16 (312 文), 1/4 (1,250 文), 1/1 (5,000 文) と変更した場合の学習曲線を

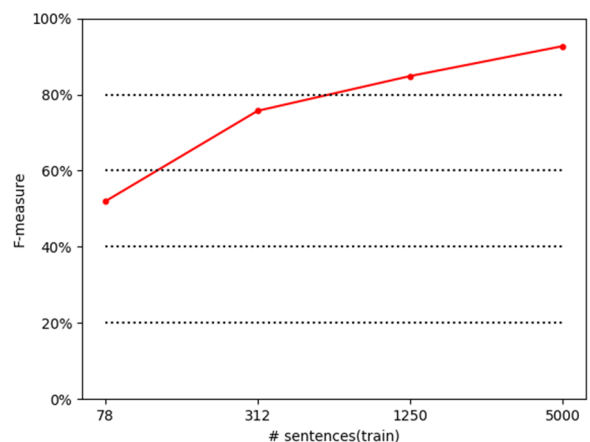


図 5 BiLSTM-CRF の学習曲線
Figure 5 Learning curve of BiLSTM-CRF

原文: 福岡、熊本、佐賀 3 県の漁業団体は 1 日、国営諫早湾干拓事業（長崎県）を巡って… 時空間表現の抽出: 福岡/Loc-B、/O 熊本/Loc-B、/O 佐賀/Loc-B 3/Loc-B 県/Loc-I の/O 漁業/O 団体/O は/O 1/Tim-B 日/Tim-I、/O 国営/O 諫早/Loc-B 湾/Loc-I 干拓/O 事業/O (/O 長崎/Loc-B 県/Loc-I) /O を/O 巡/O つ/O て/O … ※上記で Loc は文字列を地名として、Tim は時間として抽出したことを示している。 知識ベースとの照合: …'熊本': [{'geonameid': '1858419', …}], …… '1 日': [{'hutime': '/1001.1/date/平成 30 年 5 月 1 日', …}], …
--

図 6 時空間表現の正解例

Figure 6 Correct answer example of spatiotemporal expression

描画した(図 5)。図の学習曲線から、現有の最大の学習データに対応する右端の点でも右肩上がりだと言える。したがって、学習データの追加により精度向上が見込まれるといえる。

実際に、上記の推定実験により学習した時空間表現認識器 BiLSTM-CRF に実際の新聞記事を与えた結果の例を図 6 に示す。空間表現である「福岡」や「熊本」等の地名を正しく認識できており、さらに「1 日」も時間表現として認識できていることが図から確認できる。しかしながら、一般の地域研究のテキストを対象とすることを考えると、特に空間表現は対象とする地域の表現が頻出すると想定される。高い精度が要求される場合には、同ドメインの学習データを用意し、モデルを再学習させる必要があるといえる。

なお、新聞記事データにおける 1 記事あたりの計算時間は、Core i7 6850K (3.6GHz) の 1 コアを用いたとき、単語分割 (KyTea) は 0.04[s]、時空間表現認識 (BiLSTM-CRF) は 0.49[s] だった。このことから時空間表現認識が律速段階であるといえる。両処理とも文単位で独立であるので、並列化は極めて容易であり、大量のテキストであっても計算時間の問題は起きにくい。

テキスト中の時空間表現が認識されると、時間表現は絶対時間（国際規格 ISO8601 を元に曖昧

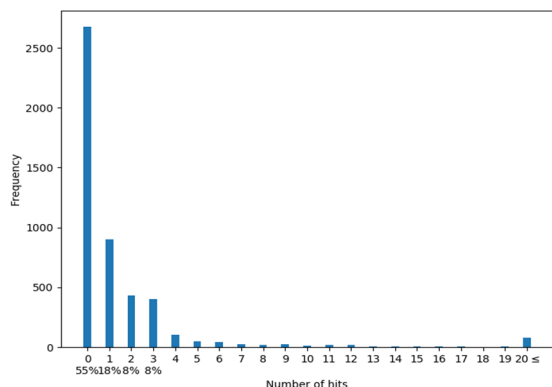


図 7 空間表現の知識ベースにおけるヒット数

Figure 7 Number of hits in knowledge base of spatial representation

性を許容するように拡張した表現形式) に、空間表現は緯度経度の組 (複数の可能性あり) を付与する。我々は時間と空間の知識ベースを有しており [8, 9], 個々の表現を知識ベースのエントリーに対応付けることで緯度経度の組を付与した。したがって、知識ベースに記述がない空間表現や複数のエントリーがある空間表現が存在することになる。知識ベースとの照合により得られたエントリー数を調査した結果を図 7 に示す。

2.3 新聞記事からの話題検出

ここでは新聞記事における話題 (トピック) を自動的に検出するための手法について述べる。本研究ではトピック検出手法としてトピックモデルの 1 つである LDA (Latent Dirichlet Allocation) [10] を用いた。LDA は統計的に共起しやすい用語の集合をトピックとして表現していく。

2.3.1 用語抽出

LDA ではトピック検出の対象を Bag-of-Words (BOW) で表現する必要がある。そこで、2.2 節の手法により新聞記事本文を単語分割し、この結果から新聞記事の特徴づける用語を抽出する。具体的には、名詞および名詞の連続を用語の対象とした。名詞もしくは名詞の連続の直後に接尾辞がある場合、接尾辞も含めて用語として抽出した。抽出した用語およびその新聞記事ごとの出現頻度を組み合わせることで BOW を作成した。

2.3.2 トピックモデルの適用

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新聞記事ごとのトピックの分布をモデル化する。図 8 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 は新聞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 β 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。新聞記事数が D 、新聞記事 d の用語数が N_d であるとき、

$$\theta_d \sim \text{Dir}(\alpha) \quad (d = 1, 2, \dots, D),$$

$$\phi_k \sim \text{Dir}(\beta) \quad (k = 1, 2, \dots, K)$$

により生成されると仮定する。ここで $\text{Dir}(\cdot)$ はデ

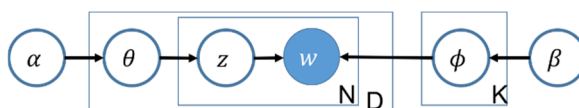


図 8 LDA のグラフィカルモデル

Figure 8 Graphical model for LDA

イリクレ分布を示す．トピック $z_{d,i}$ は $Multi(\theta_d)$ ($i = 1, 2, \dots, N_d$) により生成される．ここで $Multi(\cdot)$ は多項分布を示す．用語 $w_{d,i}$ は

$Multi(\phi_{z_{d,i}})$ により生成すると仮定できる．LDA のモデル推定では崩壊型ギブスサンプラを用いた解法が知られており[11]，本研究でもそれも用いてトピックを検出した．

2.3.3 学習データを用いたトピック検出

先行研究[12]において，2010 年から 2015 年の 6 年間の毎日新聞記事 (CD-毎日新聞 2010～2015 データ集[13]，新聞記事数：606,924 件，用語の異なり数：2,683,289 語，延べ数：286,288,248 語) に対して LDA によるトピック検出を実施した．この検出結果を学習データとして用いることで，新たに収集した新聞記事からのトピック検出を試みる．

ここで，学習に利用した 2010 年から 2015 年までの毎日新聞記事集合を D ，2.1 節により収集した新聞記事集合を F ，サンプリング回数を S とした場合の崩壊型ギブスサンプリングの手続きを図 9 に示す．この手続きでは学習データに対する再度のサンプリングは行わず，新たに収集した新聞記事のみを対象とした．これにより，学習データにおけるトピック検出の結果自体は変化することなく，学習データの特徴に基づき，新データのトピックを検出することができる．

2.3.4 実験

収集した新聞記事を対象にトピック検出を行った．LDA における K (トピック数) を 200 とし，図 9 におけるギブスサンプリングを 2,000 回繰り返した．対象とした新聞記事は 30,753 件，用語の異なり数は 165,994 語，延べの用語数は 1,870,598 語であった．例えば，図 6 の新聞記事の

1. Initialize α and β
2. Initialize z
3. Set S : the number of sampling
4. for $s = 1, \dots, S$ do
5. for $d = D+1, \dots, D + F$ do
6. for $i = 1, \dots, N_d$ do
7. Sample $z_{d,i}$
8. Update $N_{d,z_{d,i}}$
9. end for
10. end for
11. end for

図 9 学習データを用いた崩壊型ギブスサンプリングの手続き

Figure 9 Procedure of collapsed Gibbs sampling with learning data

場合，出現する語彙とトピックの関係は，福岡/34，熊本/167，佐賀 3 県/105，漁業団体/105，国営諫早湾干拓事業/105 のようになった（ここで/の後ろの数字がトピックを示している）．なお，1 新聞記事あたりの計算時間は，Core i7 (4.0GHz) の 1 コアを用いたとき，おおよそ 1.17[s]であった．またトピック数の 200 であるが，これはトピック数を 30, 50, 100, 150, 200, 250, 300 としたトピック検出を予備的に実施し，トピック数と検出内容を比較した結果から 200 が適当であると判断した．

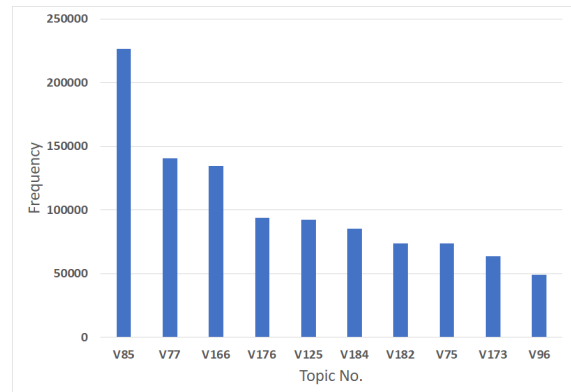


図 10 LDA によるトピック検出結果
Figure 10 Topic detection by LDA

図 10 は割り当てられた用語数順に上位 10 件までのトピックを表示している．そのうち上位 5 件のトピックと割り当てられた用語は次のとおりである．

- トピック 85 : こと，ため，もの，前，...
- トピック 77 : 男性，女性，逮捕，県警，事件，男，...
- トピック 166 : 問題，説明，調査，対応，...
- トピック 176 : 米国，ロシア，イラン，イスラエル，...
- トピック 125 : 中国，日本，会談，米国，協議，表明，合意，北朝鮮，...

これら以外にも，国会関連（トピック 184），株式市場関連（トピック 159），オリンピック関連（トピック 75），大震災関連（トピック 3），災害関連（トピック 117）などがあつた．トピック 85 や 166 のように何のイベントであるかが判断できないトピックもあるが，該当期間に発生したイベントに応じたトピックは概ね検出されていると判断できる．

また，本研究の LDA では学習データを用いたトピック検出であることから，学習データとして用いた新聞記事の集合 D より検出したトピックと収集した新聞記事の集合 F より検出したトピックの対比も可能である．例えば，トピック 75 について見てみると，

- *D*: 優勝, 日本, 出場, 女子, 選手, 男子, 決勝, 大会, 五輪, ...
 - *F*: 獲得, 出場, 優勝, 金メダル, メダル, 2位, 決勝, 五輪, ...
とあまり変化がない. 一方で, 大相撲関連 (トピック 116) では
 - *D*: 1, 横綱, 白鵬, 2, 0, 大関, 押し出し, 里, 相撲, 3, 鶴竜, 土俵, 稀勢, 日馬富士, ...
 - *F*: 元横綱, 相撲, 貴乃花親方, 日本相撲協会, 横綱, 協会, 白鵬, 土俵, 日馬富士, 鶴竜, ...
- のように, *D*では大相撲の取り組みに関するトピックと解釈できるが, *F*では大相撲ではあっても貴乃花の騒動にトピックが移っている印象を受け, 少なからず趣が異なっている.

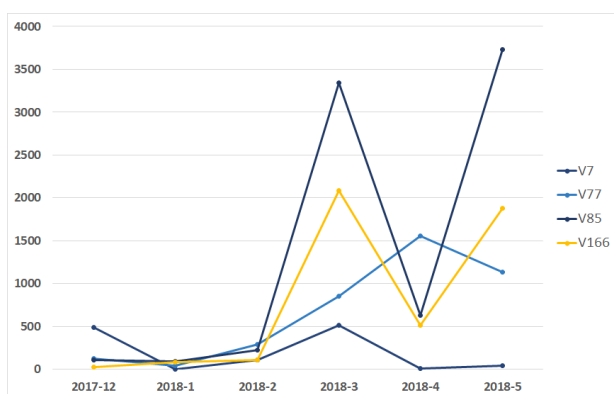


図 11 愛媛県に関するトピックの推移
Figure 11 Transition of topics about "Ehime"

2.2 節の時空間表現の抽出により, 地名として異なり数 4,838 語, 異なり数 74,644 語を得ることができた. この抽出した地名ごとに検出したトピックを割り当てることで, 地名ごとのトピックを得ることができる. 図 11 は「愛媛県」, 「松山市」, 「今治市」に関するトピックの月単位の時

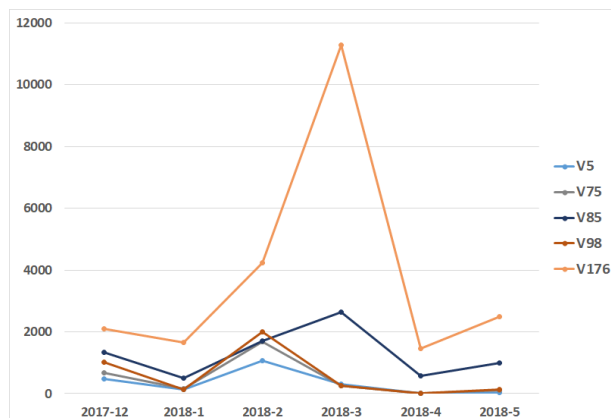


図 12 ロシアに関するトピックの推移
Figure 12 Transition of topics about "Russia"

系列変化を示す. 2018 年 3 月以降に新聞記事数が増えており, 特にトピック 85 と 166 の頻度が高くなっている. トピック 85 には「財務省」, 「森友学園」, 「学校法人」といった用語が割り当てられているためだと考えられる. また, トピック 7 の出現も見受けられる. これは原発に関するトピックである.

図 12 は「ロシア」に関するトピックである. 当然のことながら, トピック 85 は出現するものの, 全体や他の地名と比較すると頻度は低かった. 特に目立つのはトピック 176 だった. また, 2018 年 2 月にはトピック 98 (オリンピック全般), 75 (オリンピックの試合結果関連), 5 (フィギュアスケート関連) が他の月に比べ多かった

2.4 時空間情報の可視化

2.2 節で示したように, 時空間情報認識処理の過程で各新聞記事には時間と空間の知識ベースへのリンク情報が付与されているので, これを利用すれば, 新聞記事を地図やタイムライン上に配

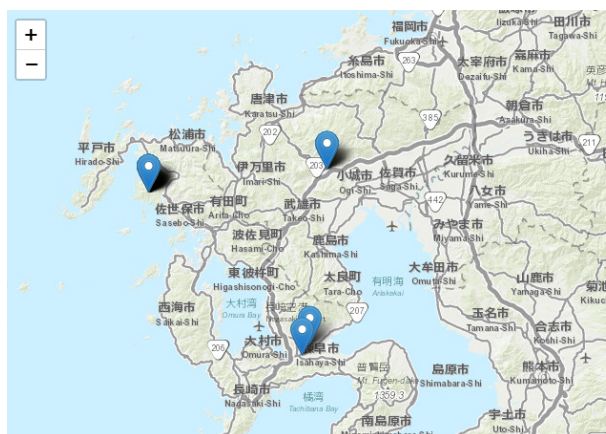


図 13 可視化の例
Figure 13 Example of visualization

- 2017-12-05 諫早湾干拓:高裁で和解協議再開へ 請求棄却訴訟控訴審
- 2017-12-18 諫早排水:ノリ色落ちで海上デモ 佐賀の漁業者
- 2018-02-01 佐賀県漁協:諫早和解も 開門要求を転換 国基金受託検討
- 2018-02-26 諫早湾訴訟:高裁で和解協議再開 国と漁業者の妥協 焦点
- 2018-02-26 諫早和解協議再開:打開へ新案提示を
- 2018-03-05 諫早訴訟:福岡高裁和解案を漁業者側拒否 7月30日判決
- 2018-03-06 諫早訴訟:漁協「踊らされた」 和解案、長崎は歓迎
- 2018-05-15 諫早訴訟控訴審:開門案を含む和解協議求める要望書提出
- 2018-05-22 諫早・鳥害訴訟:長崎県の漁業者4人、補助参加申し立て
- 2018-05-28 諫早訴訟:和解協議は決裂 福岡高裁「開門」無効化判決へ

置できる。このような可視化により、新聞記事に付随するトピックを介して、地域の特性を把握することが可能になるのではないかと考えている。本稿のタイトルである「Web ビッグデータからの地域研究情報抽出の試み」に関わるツール化の事例である。

そのような発想のもとでの試作ツールを図 13 に示す。これは、2.2 節および 2.3 節で作成されたデータを json 形式で HTML 文書に取り込むことで、Web アプリケーションとして可視化を実現したものである。

この例では、特定のトピック (2.3 節で V105 などの ID が付けられ、「開門」、「漁業者」、「福岡高裁」、「諫干」といった単語に特徴づけられているトピック) の記事をマップ上に並べ、記事のリストを時系列順に並べている。またピンのそれぞれをクリックすることで、対応する記事の確認もできる。このインタフェースによって、個々のニュースを確認しながら、あるトピックに関わる動向の全体像を把握することができる。

3. 考察

研究手法として Web から収集した新聞記事进行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うえて Social Media Site 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科学研究の事例もある。これらの研究では、対象とする情報源を限定した上で、あらかじめ設定したキーワードにより記事を収集している。研究対象が定まっている場合において、この手法は有効である。

しかし、Web 上のビッグデータを利用して、地域の動向を包括的に俯瞰したり、通常と異なる状態を検出したりすることが目的の場合、キーワードを事前に設定することはできない。本稿の課題である「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して可視化する情報システムの開発」は、この事例に対応する。そのような情報システムの構築をめざし、本研究では Web 上の新聞記事を対象として、①Web Scraping によるデータ収集、②BiLSTM-CRF に基づくニューラルネットワークを用いた語彙の抽出と時空間表現の認識、③トピックモデルによる話題抽出、④可視化という情報処理技術の組み合わせを試みた。新聞記事を対象と利用した理由は、データ収集や分析の容易さによる。本稿執筆の時点において、上記の各部分は個別に処理されているが、速やかに統合する予定である。

Web Scraping によるデータ収集は大きな問題も無く稼働している。ただし、新聞社の多くが新聞記事の有料化を進めているため、何らかの方策を講ずる必要がある。

BiLSTM-CRF に基づくニューラルネットワークを用いた語彙の抽出と時空間表現の認識につ

いては、テストデータの都合により新聞記事を対象とした認識精度は計測していない。しかし、テストデータによる実験から得られた学習曲線 (図 5) およびテストデータと新聞記事の構造は大きく変わらないことから、本法の有効性については目処がたったと考えている。ただし、実用化に向けて、さらに以下の課題が挙げられる。

1. エントリーがない空間表現に対して周囲の空間表現などから緯度経度を推定する。
2. エントリーが複数ある空間表現に対しては周囲の空間表現などから 1 つに確定する。

例えば、前者は図 6 における「諫早湾」のエントリーが存在しない場合に、「福岡」や「熊本」等の周囲の空間表現から緯度経度を推定できることに相当する。また後者に関しては、「熊本」という地名に対して、「熊本県」だけでなく、福岡の嘉麻市や北九州市や田川郡にみられる「熊本」もエントリーとして存在する場合に、周辺の時空間表現を参照して一意に決定できることに相当する。

トピックモデルによる話題抽出も有効性については目処がたったと考えている。本研究では、過去の一定期間の学習データに対して LDA によるトピック検出を実施し、その検出結果を新たに収集した新聞記事からのトピック検出に利用している。これは学習データと新しいデータのトピック検出に一貫性を持たせるためである。しかし、学習データと新しいデータの時間差が大きくなるにつれて、同じトピックであっても内容に変化が生ずるため (2.3.4)、定期的な再学習が必要になると考えている。そのため、トピックの分化、統合、さらに生成などの新たな手法が不可欠であると考えている。また今回の実験では日本語を対象としていたが、地域研究用のシステムが目的であるため、多言語への対応は不可欠である。LDA の処理は語彙に付与されている ID を用いて行えるため、言語の違いは問題とならない。つまり、BOW を作成する (2.3.1) 前段階である形態素解析からの出力に対して、辞書を用いて言語間の対応付けを行えばよい。LDA の処理は出現する語彙に依存するので同義語 (例えば計算機とコンピュータ) という問題も発生する。これについても、シソーラスを用いれば、多言語と同じ枠組みで対応できると考えている。

可視化について、現時点では単純な可視化のみであるが、GIS ツールと組み合わせることで複数のトピックをレイヤとして重ね合わせてトピック間の空間的な関連を調べるなど、発展の余地があると考えている。またトピックモデルでは 1 つの新聞記事に複数のトピックが含まれているため、トピックの構成を色調で表現するなど、トピックモデルならではの可視化ツールの構築も可能であると考えている [14]。今後、どのような可視化や分析法が地域研究者の気づきを支援できるか、インタラクティブな可視化も含めての検討も必要である。

本研究の目標は地域研究への適用であるため、対象言語を日本語から英語やスペイン語等へ拡大するとともに、データ源として Social Media Site を加える予定である。また、本研究で開発するツール群は Web 上のビッグデータから研究のヒントとなる情報を抽出するものであるが、将来的には時空間分析や異常・変換検知等の機能を用意したいと考えている。

謝辞

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バ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] The SQLite project: SQLite Home Page, <https://www.sqlite.org/index.html>
- [3] Ariya Hidayat: PhantomJS, <https://github.com/ariya/phantomjs>
- [4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton: Speech Recognition with Deep Recurrent Neural Networks, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (2013).
- [5] Huang, Zhiheng, Wei Xu, and Kai Yu: Bidirectional LSTM-CRF Models for Sequence Tagging, arXivpreprint arXiv:1508.01991 (2015).
- [6] 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.
- [7] Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den: Balanced corpus of contemporary written Japanese, Language resources and evaluation 48.2 (2014): 345-371.
- [8] HuTime: <http://www.hutime.jp/>
- [9] Shoichiro Hara: Digital gazetteer as a knowledgebase for open data science, DOI: 10.23919/PNC.2017.8203524
- [10] D.M.Blei, A.Y.Ng, and M.I.Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3, pp.993-1022(2003).
- [11] T.L.Griffiths and M.Steyvers: Finding scientific topics, Proc. of the National Academy of Sciences of the United States of America, vol.101, pp.5228-5235(2004).
- [12] 山田太造: 新聞記事に対するトピックモデルの適用とトピックの時系列変化に関する考察. 研究報告人文科学とコンピュータ (CH), Vol.2017-CH-115, No.1, pp.1-5, 2017.
- [13] CD-毎日新聞 2010~2015 データ集: <http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html>
- [14] 高田百合奈・渡邊英徳・柳澤雅之・山田太造: 位置情報とトピックモデルに基づくフィールドノートのビジュアライズ手法, じんもんこん 2014 論文集, 3 号, 2014 年, 57-62 頁, <http://id.nii.ac.jp/1001/00107362/>