

# Prototyping Information System to Extract Area Study Information from Web Big Data

Hara, S.,<sup>1</sup> Yamada, T.,<sup>2</sup> Ishikawa, M.,<sup>3</sup> Shirai, K.,<sup>4</sup> Kameda, A.<sup>5</sup> and Mori, S.<sup>6</sup>

<sup>1</sup>Center for Southeast Asian Studies, Kyoto University, Japan

<sup>2</sup>Historiographical Institute, University of Tokyo, Japan

<sup>3</sup>Faculty of Business Administration, Tokyo Seitoku University, Japan

<sup>4</sup>Graduate School of Informatics, Kyoto University, Japan

<sup>5</sup>Center for Southeast Asian Studies, Kyoto University, Japan

<sup>6</sup>Academic Center for Computing and Media Studies, Kyoto University, Japan

## Abstract

*With the spread of Internet, abundant information related to area studies has been distributed on the Web. Even if the information is limited to text media, the amount of data available on the Web is still enormous. Thus, it is more difficult nowadays to extract research relevant information, and it is impossible for researchers to read and analyze all the extracted data as has been done before. This paper will propose a solution to this drawback through the development of information system that automatically extracts, analyzes and allows the visualization of bigdata on the Web. This paper explores new directions for area studies based on informatics that is compatible with the Internet age.*

## 1. Preface

Area studies cover inter-disciplinary areas encompassing humanities, social sciences, natural sciences, engineering, health and medicine, etc. (Hara, 2010). Focusing on humanities, the main research resources used by researchers until now have been text-based media, such as historical records, literary works, research papers, newspapers, and magazines. Area researchers have been engaged in analyzing the texts' contents during the process of research. However, with the spread of the Internet, multiple types and large volumes of information related to area studies are being distributed on the Web. This trend has become even more prominent with the occurrence of large-scale disasters, political conflicts or changes, national elections, etc. Even if limited to text media, there is a huge volume of data on the Web, and it has become impossible for humans to read and analyze all the extracted data.

Therefore, this study attempts to develop an information system that automatically extracts, analyzes and visualizes bigdata on the Web, which has the potential to provide some hints to start new researches. This study also allows for the exploration of new directions for area studies based on informatics that is compatible with the Internet age.

## 2. Method and Results

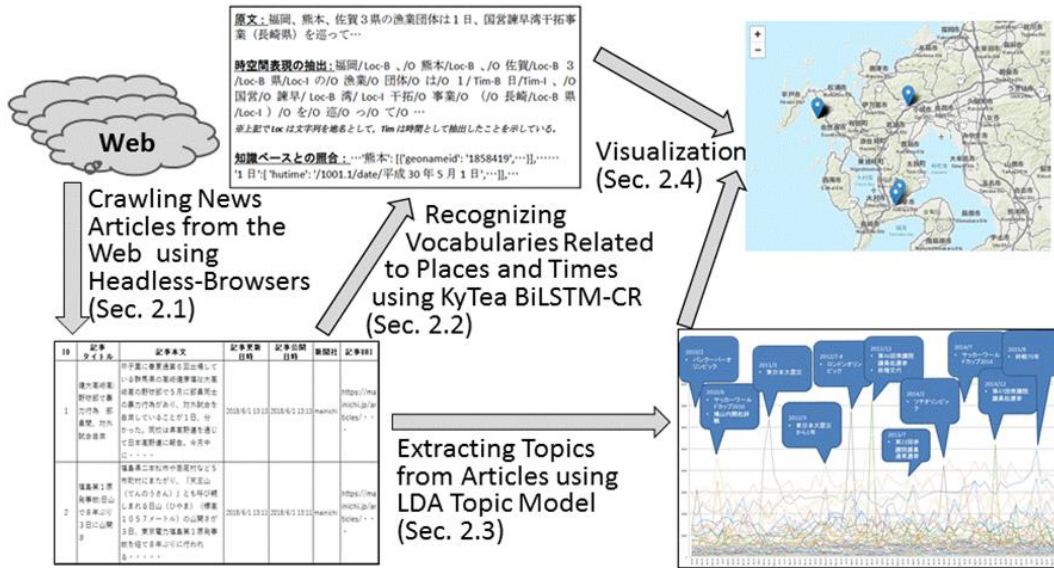
This study uses Internet news articles as experimental sources in extracting useful area study information from Web Big data. The reasons for this

choice are that, (1) compared to social media sites, such as Twitter or Facebook, the texts included in newspapers are more consistent in terms of the use of grammar and vocabularies, which makes it easier to apply natural language processing; (2) newspapers are useful information sources for comprehensively gaining information on topics such as politics, economics, and culture; and (3) most of the newspaper's companies publish articles not only in paper media, but also as digital data on the Web, which allows for efficient data collection and processing.

In this study, following the methods shown in Figure1, the research begins with the collection of newspaper articles from the Web (explained in section 2.1), followed by the recognition the words specially related to spatiotemporal information in newspaper articles (section 2.2), and extracts themes from the newspaper' contents (section 2.3), which finally allow for the spatial visualization of automatically selected relevant themes (section 2.4).

### 2.1 Collecting Newspaper Articles from the Web

Locations in which newspaper articles are published on the Web include not only newspaper companies' own sites but portal sites and curation software. In order to collect newspaper articles from multiple companies, portal sites are more efficient. However, the newspaper articles published on a portal site only present a part of entire set of articles published on a newspaper company's site.



Usually, a portal site manages the process of selecting and forwarding an article from a newspaper site into another Web page. Therefore, this study has envisaged a specific module which allows the collection of newspaper articles directly from the newspaper companies' own site (hereafter, "this newspaper article collection module"). However, as most of the newspaper companies are starting to charge for some of the articles, this study only covers the portion of articles that can be viewed free of charge. Thus, this study collects articles from Mainichi Shimbun, Asahi Shimbun, Yomiuri Shimbun, and AFP (L'Agence France-Presse: English language version). This newspaper article collection module gathered the following information from the newspaper company sites.

- Article titles
- Article main text
- Revision date/time
- Published date/time (if an article does not contain published date/time, the revision date/time is used. See (3) below)
- Article URL (URL for the Web page on which the main text of the newspaper article is published)

In addition to the above, this newspaper article collection module records the data management system ID (a simple serial number) and the newspaper company name in order to identify the newspaper article provider. The procedure for collecting newspaper articles is follows:

- (1) Acquire a revision list of newspaper articles from the newspaper companies' site. A revision list refers to a collection of attributes

of a newspaper article such as article title, revision date/time, and article's URL. Revision lists of the Mainichi Shimbun and the Asahi Shimbun are collected using RSS (RDF Site Summary). However, as the Yomiuri Shimbun and AFP do not provide RSS, the revision lists of these newspapers are compiled using relevant data extracted from Web pages which include the descriptions about newly-arriving newspaper articles. An example of the newspaper article is shown in Figure 2.

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と. 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>

Figure 2: An example of aggregated article

- (2) Confirm if a newspaper article has already been collected or not. Extract the first attributes' set from the revision list. The article title and revision date/time from the extracted attributes' set are compared with the attributes' set included in the database created by this newspaper article collection module. If the same attributes' set is absent from the database, the article is a newly arriving one and procedure (3) is executed. If the same attributes' set exists in the database, the article has already been collected and the next attributes' set in the revision list is processed.

- (3) Collect the newspaper article. Based on procedure (2), if the newspaper article is a newly arriving, download the Web page including the main text of the newspaper article from the newspaper company site by referencing to the URL in the revision list. Next, extract HTML elements corresponding to the main text of the newspaper article from the downloaded Web page. Note that the newspaper articles on the Web sometimes include subheadings within the main texts in addition to the main heading at the top of the article. However, on the contrary to the Web newspaper articles, subtitles rarely appear in paper version articles. Therefore, subheadings within main texts of web newspaper articles were not collected in this study. Additionally, in the case of the Mainichi Shimbun and the Yomiuri Shimbun, the publishing date/time of the newspaper articles are described in the meta elements of the main texts of the Web page. Thus, the publishing date/time can be collected simultaneously during the process of collection of the article's main text. Finally, the data collected in procedures (2) and (3) are registered in the database of this newspaper article collection module as a new newspaper article. Procedures (2) and (3) are repeated until there are no more attributes' set in the revision list.
- (4) In consideration to the access burden that this collecting module may infringe to the newspaper company sites which host the articles to be collected, an appropriate interval time is envisaged before repeating procedures (1) to (3). This preventive action enables newspaper articles to be continuously collected.

This newspaper article collection module converts the collected newspaper articles into a CSV format text file for each newspaper company, and automatically processes this file following the next steps of the process (section 2.3). The character encoding for this text file is UTF-8 (without BOM). An example of the output is shown in Figure 3. This newspaper article collection module is written in Python 3.6, and, as its main libraries, this module uses feed parser for analyzing the structure of RSS, and BeautifulSoup4 for analyzing HTML documents. For the database that manages the collected newspaper articles, we use SQLite3 (The SQLite project).

In recent years, acquiring Web pages from websites has become more difficult. This is because in order to obtain relevant web pages, the number of cases that requires pre-processing interventions with cookies or JavaScript, among others, has increased. The same is true for newspaper companies' sites. As a means for solving this kind of problem, this newspaper article collection module uses a headless browser, which is widely used for crawling, to obtain the Web page rendering results from the website, in order to acquire the main text of the newspaper article. This newspaper article collection module uses PhantomJS (Ariya Hidayat) as the headless browser.

The number of newspaper articles collected in this newspaper article collection module comprised 24,851 articles from the Mainichi Shimbun (from December 2017 to September 2018), 13,903 articles from the Asahi Shimbun (from June 2018 to September 2018), 10,552 articles from the Yomiuri Shimbun (from June 2018 to September 2018), and 12,137 articles from AFP (from June 2018 to September 2018).

ID	Title of Article	Main Text of Article	Revision Date	Publishing Date	Newspaper Company	Article URI
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、国営諫干湾干拓事業（長崎県）を巡って堤防開門を強硬しないうちで、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発表した。これまで開門を求めてきた……	2018/5/1 21:38	2018/5/1 21:38	mainichi	<a href="https://mainichi.jp/articles/20180501/k00/00e/020/34500c">https://mainichi.jp/articles/20180501/k00/00e/020/34500c</a>
243	三井住友と大和証券：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（FG）と大和証券グループ本社は、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FGが60%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	<a href="https://mainichi.jp/articles/20180501/k00/00e/020/34501c">https://mainichi.jp/articles/20180501/k00/00e/020/34501c</a>

Figure 3: Output example of aggregated articles in CSV format

Although there were uncollected newspaper articles due to equipment malfunctions or changes to the newspaper company site specifications, in general, the newspaper articles have been continuously collected during the above cited period.

## 2.2 Recognition of Spatiotemporal Information from Newspaper Articles

A machine-learning-based system was developed to recognize spatiotemporal expressions from newspaper articles collected from the Web using the procedure described in the previous section. More specifically, at first, the sentences comprised in a newspaper article are segmented into words. Then, using neural network, each word is identified as either being a word in a time expression (see Figure 4, indicated by tag Tim) or a word in a spatial expression (same, indicated by Loc). Since these expressions are a sequence of words, words in a time or spatial expression are annotated with “B” indicating the starting word of the expression or “I” indicating the subsequent words in the expression simultaneously. For example, Loc-B means the first word of a spatial expression.

The neural network adopted here is BiLSTM-CRF (Huang et al., 2015) which is a combination of a Bi-directional Long Short Term Memory (BiLSTM) network (Graves et al., 2013) and Conditional Random Fields (CRF). BiLSTM-CRF is known for achieving high precision in the recognition of named entity tasks. Additionally, as the characters are used as BiLSTM-CRF processing units, in this study, it is assumed that the input characters are segmented into words using a morphological analyzer KyTea (Neubig et al., 2011), and that BiLSTM-CRF uses the words as processing units. The general diagram of the BiLSTM-CRF model is shown in Figure 4.

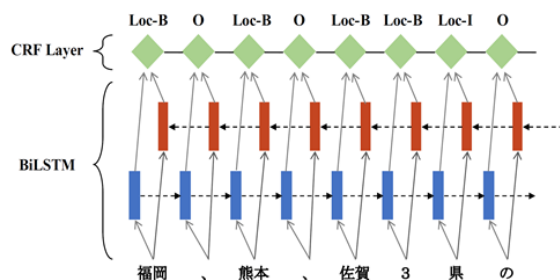


Figure 4: Schema of BiLSTM-CRF model

Here, the rectangles indicate BiLSTM (the bottom side is LSTM loading the word string from start to finish, and the top side LSTM is loading in the reverse direction). The green trapezoids express the CRF layer. BiLSTM-CRF takes word strings as

input, and estimates a corresponding BIO tag (Tim-B, Tim-I, Loc-B, Loc-I, O) for each word.

The morphological analyzer KyTea has been trained on the core data (557,281 sentences, segmented into words annotated with part-of-speech) of the “Balanced Corpus of Contemporary Written Japanese” (BCCWJ) (Maekawa et al., 2014), and it has 98% or greater accuracy on a test data from the same domain (3,024 sentences).

The parameters of the spatiotemporal expression recognizer BiLSTM-CRF are estimated using the training data selected from BCCWJ white paper (OW) whose BIO tags are attached to each word manually. In this estimation experiment, the BiLSTM-CRF recognition accuracy -- after training from 5,000 sentences used as training data -- was approximately 90% out of 389 sentences of the test data according to the recognition accuracy of the spatial expression (F value). This recognition accuracy of 90% demonstrates the effectiveness of BiLSTM-CRF. However, from a practical perspective, this might still be insufficient. Additionally, we can assume that the accuracy may change due to differences between the training data BCCWJ (OW) and the newspaper article domains to which it is applied.

To investigate the possibility of improving accuracy by adding training data, we drew a learning curve (Figure 5). The curve shows the change in an accuracy rate in relation to the volume of training data incorporated.

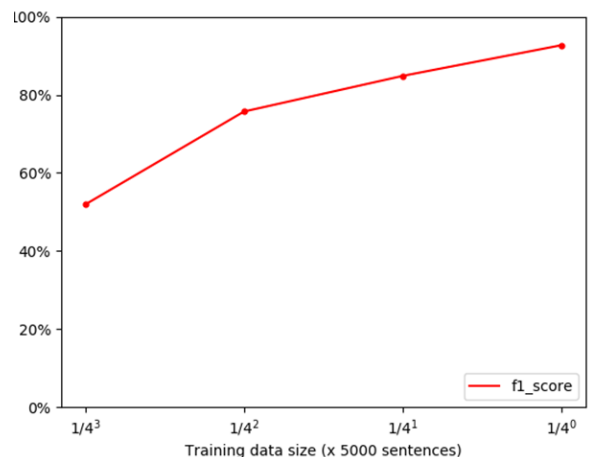


Figure 5: Learning curve of BiLSTM-CRF

In Figure 5, the training data changed from 78 sentences (1/64 of the total 5,000 sentences of the training data), 312 sentences (1/16), 1,250 sentences (1/4) to 5,000 sentences (1/1, i.e. all training data). The learning curve in Figure 5 demonstrates that even the points on the right edge are on an upward progression in relation to the current maximum

training data. Therefore, it is possible to forecast that accuracy will improve with an increase in training data.

Figure 6 shows an example of the actual results of applying real newspaper articles to the spatiotemporal expression recognizer BiLSTM-CRF learned from the estimation experiment described above.

**Original Text:** 福岡、熊本、佐賀3県の漁業団体は1日、国営諫早干拓事業(長崎県)を巡って...

**Extracted Spatiotemporal Expressions:** 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/Time-B 日/Time-I、/O 国営/O 諫早/Loc-B 干拓/O 事業/O (/O 長崎/Loc-B 県/Loc-I )/O を/O 巡/O っ/O て/O ...

**Linked with Knowledge Databases:** ...'熊本':[['geonameid': '1858419', ...]], ...'1日':[['hutime': '/1001.1/date/平成30年5月1日', ...]], ...

Figure 6: Correct answer example of spatiotemporal expression using BiLSTM-CRF

This figure confirms that BiLSTM-CRF correctly recognizes the spatial expressions “福岡 (Fukuoka)” and “熊本 (Kumamoto),” as much as, it recognizes “一日 (one day)” as a temporal expression. However, as the target of this study is area studies, spatial representations of a particular area will appear frequently. In cases when high accuracy is required, it will be necessary to prepare specific training data for the same domain and allow the model the possibility to re-learn.

The calculation time per newspaper article data, when using a single Core i7 6850K (3.6GHz), was 0.04 [s] for the word segmentator (KyTea), and 0.49[s] for the spatiotemporal expression recognizer (BiLSTM-CRF). This fact allows saying that spatiotemporal recognition is at the rate-limiting stage. The word units are independent for both processes, so parallelization is very simple, and even when there are large volumes of text, there will unlikely occur any issues related to calculation time. When spatiotemporal expressions are recognized in texts, the temporal expressions are converted to absolute times (expression format enhanced to allow ambiguity based on international standard ISO8601), and the spatial expressions are converted to sets of longitude and latitude (multiple sets are also possible).

Previous research of this team has already developed spatiotemporal knowledge bases which register pairs of spatiotemporal expressions and their values (time, or longitude and latitude) (HuTime and Hara, 2017), thus spatiotemporal expressions are converted into values that can be

searched from the already built knowledge base. However, some place names have not been registered in the knowledge bases, but some place names might be registered more than one time. Figure 7 counts place names by the number of entries (or hits) in the spatial knowledge database (e.g., there are about 400 place names which have 3 entries in the database). This shows the fact that many spatial expressions are not registered in the knowledge database. This problem will be discussed in section 3.

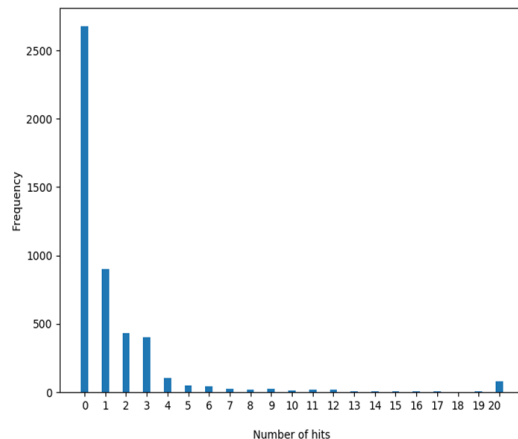


Figure 7: Number of hits in knowledge base of spatial expressions

### 2.3 Detecting Topics from Newspaper Articles

A method to automatically detect topics in newspaper articles was developed. In this study, LDA (Latent Dirichlet Allocation) (Blei et al., 2003), which is called as “topic model,” was used as a topic detection method. LDA expresses the sets of words that are statistically likely to co-occur as a topic.

#### 2.3.1 Extracting words

With LDA, it is necessary to express detected topics as Bag-of-Words (BOW). Therefore, firstly, newspaper article main texts are segmented into words using the method described in section 2.2, and then some words that characterize a newspaper article are extracted. In concrete terms, nouns or a series of nouns are the targets of words. In case there is a suffix immediately after a noun or a series of nouns, this is extracted as a word that includes the suffix. The BOW is created as a combination of extracted words and their frequency of expression for each newspaper article.

#### 2.3.2 Application of topic models

The LDA assumes that multiple topics exist within one document, therefore, this method models both

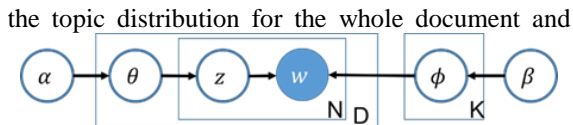


Figure 8 Graphical model for LDA

Figure 8 shows the LDA graphical model used in this study. Here, the blue circle represents the observed variables, while the white circles represent unknown variables. The rectangles represent repetitions, and a symbol on the bottom right in each rectangle shows the number of repetitions. Here,  $w$  is the only observed variable which indicates the extracted word.  $z$  is the topic,  $\theta$  is the topic distribution for each newspaper article, and  $\phi$  indicates word distribution for each topic.  $\alpha$  and  $\beta$  are hyperparameters for  $\theta$  and  $\phi$  respectively. If the number of newspaper articles is  $D$ , the number of topics is  $K$ , and the number of words from newspaper articles  $d$  is  $N_d$ , it is supposed that:

$$\begin{aligned}\theta_d &\sim \text{Dir}(\alpha) \quad (d = 1, 2, \dots, D), \\ \phi_k &\sim \text{Dir}(\beta) \quad (k = 1, 2, \dots, K)\end{aligned}$$

Equation 1

are generated. Here,  $\text{Dir}(\cdot)$  indicates Dirichlet Distribution. Topic  $z_{d,i}$  (which means the topic from which  $i$ -th word in document  $d$  comes) is generated from:

$$\text{Multi}(\theta_d) \quad (i = 1, 2, \dots, N_d)$$

Equation 2

Where,  $\text{Multi}(\cdot)$  indicates Multinomial Distribution. Terms  $w_{d,i}$  (which means  $i$ -th word in document  $d$ ) can be supposed to be generated from:

$$\text{Multi}(\phi_{z_{d,i}})$$

Equation 3

As Collapsed Gibbs Sampler is a well-known solution to LDA model estimation (Griffiths and Steyvers, 2004), this study also used it as well to detect topics.

### 2.3.3 Topic detection using training data

The preceding research (Yamada, 2017) detected topics by applying LDA to the Mainichi Shimbun newspaper articles published during six years from 2010 to 2015 (CD–Mainichi Shimbun (CD, 2010–2015): number of newspaper articles (606,924), number of different terms (2,683,289), total number

the topic distribution for each newspaper article. of words (286,288,248)). Using the detected results as training data, this study attempted to detect topics from the newly collected newspaper articles. The procedures for the Collapsed Gibbs Sampler in this study is shown in Figure 9, where,  $D$  is the Mainichi Newspaper article set from 2010 to 2015 used for training,  $F$  is the set of new newspaper articles collected in section 2.1, and  $S$  is the number of samplings.

1. Initialize  $\alpha$  and  $\beta$
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

Figure 9: Procedure of Collapsed Gibbs Sampler with training data

With these procedures, sampling is not performed again on the training data, and only the newly collected newspaper articles are covered. Thus, the actual results of topic detection remain unchanged in the training data, and the new data topics are detected based on the new data.

### 2.3.4 Experiment

Topic detection was performed over the collected newspaper articles. With  $K$  (number of topics) in LDA as 200, Gibbs Sampling in Figure 9 was repeated 2,000 times. The number of target newspaper articles was 30,753, the number of different words was 165,994, and the total number of words was 1,870,598 words. For example, in case of newspaper articles in Figure 6, the relationship between an appearing word and its topic is indicated by the number written after the diagonal bar “/” as follows: 福岡(Fukuoka)/34, 熊本(Kumamoto)/167, 佐賀(Saga) / 3, 県(prefectures)/105, 漁業団体(fishery industry organization)/105, and 県営諫早湾干拓事業(Isahaya Bay prefectural reclamation project)/105.

The calculation time per newspaper article, when using a Core i7 (4.0GHz), was approximately 1.17 [s]. Additionally, the number of topics  $K$  was defined as 200 beforehand in this experiment. This was determined by a preliminary test of evaluating the detected contents when the topic number  $K$  was changed from 30, 50, 100, 150, 200, 250 to 300. That is, when the number of topics exceeds 200, it

will be difficult to distinguish the difference of topics from evaluating the words' pattern in each

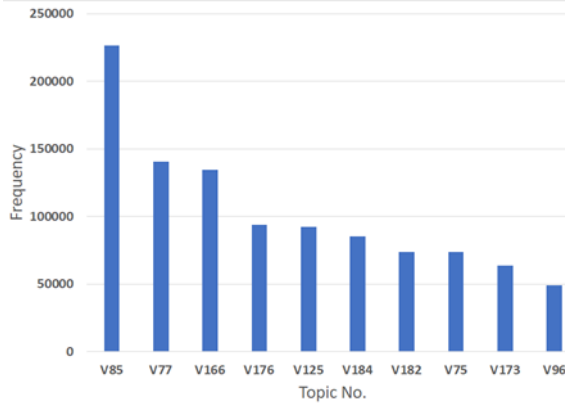


Figure 10: The top 10 topics detection by LDA

Figure 10 shows the top ten topics in terms of the number of words assigned to each topic. Of these, the frequent words assigned to each of the top five topics are as follows:

- Topic 85: こと, ため, もの, 前 ... (thing, because, object, before ...)
- Topic 77: 男性, 女性, 逮捕, 県警, 事件, 男 ... (male, female, arrest, prefectural police, case, man ...)
- Topic 166: 問題, 説明, 調査, 対応 ... (problem, explanation, survey, response ...)
- Topic 176: 米国, ロシア, イラン, イスラエル ... (U.S., Russia, Iran, Israel ...)
- Topic 125: 中国, 日本, 会談, 米国, 協議, 表明, 合意, 北朝鮮 ... (China, Japan, talk, U.S., discussion, declaration, agreement, North-Korea ...)

In addition to the words described above, there were also parliament-related words (topic 184), stock market-related words (topic 159), Olympic-related words (topic 75), major earthquake-related words (topic 3), disaster-related words (topic 117), and more. Whereas there are topics, such as topic 85 and 166, where it is not possible to judge what specific event this topic covers, this method still allows detecting topics generally related to an event that occurred during a particular period.

Additionally, as the LDA model applied in this study uses training data for topic detection, it is possible to compare topics detected from the set D of newspaper articles used for training data with the topics detected from the F of new newspaper articles collected in section 2.1. For example, there is small change in the case of topic 75:

topic.

- **Set D:** 優勝 (win), 日本 (Japan), 出場 (participate), 女子 (women), 選手 (athlete), 男子 (male), 決勝 (final), 大会 (event), オリンピック (Olympics), ...
- **Set F:** 獲得 (gain), 出場 (appearance), 優勝 (win), 金メダル (gold medal), メダル (medal), 2 位 (2nd), 決勝 (final), 五輪 (Olympics), ...

On the other hand, with a Sumo-related topic (topic116):

- **Set D:** 1 (1), 横綱 (Yokozuna), 白鳳 (Hakuho), 22(22), 0(0), 大関 (Ohzeki), 押し出し (Oshidashi), 里 (Sato), 相撲 (Sumo), 33(33), 鶴竜 (Kakuryu), 土俵 (Sumo ring), 稀勢 (Kise), 日馬富士 (Harumatafuji), ...
- **Set F:** 元横綱 (Former Yokozuna), 相撲 (Sumo), 貴乃花親方 (Takanohana master), 日本相撲協会 (Japan Sumo Association), 横綱 (Yokozuna), 協会 (association), 白鳳 (Hakuho), 土俵 (Sumo ring), 日馬富士 (Harumatafuji), 鶴竜 (Kakuryu), ...

In the example above, the topic 116 in set D can be interpreted as a general topic on Sumo matches, but the same topic in Set F gives a different impression that the topic has shifted to a dispute concerning the master Takanohana.

The extraction of spatiotemporal expressions in section 2.2 allowed obtaining 4,838 different words for place names, and 74,644 different words. Once assigning detected topics to each of these extracted place names, it is possible to obtain topics for each place name. Figure 11 shows time series changes in monthly units for topics related to “愛知県 (Aichi Prefecture)”, “松山市 (Matsuyama City)”, and “今治市 (Imabari City).” Since March 2018, the number of newspaper articles have been increasing, and, in particular, the frequency of topics 85 (V85 in the Figure 11) and 166 (V166) have been increasing. For topic 85 (V85), the probable cause was the political scandal whose related words were “財務省 (Ministry of Finance),” “森友学園 (Moritomo Academy),” and “学校法人 (incorporated school).”

Additionally, the topic 7 (V7) appears, which is a topic related to nuclear power.

### 2.4 Visualization of Spatiotemporal Information

As shown in section 2.2, as part of the process of spatiotemporal information recognition processing, link information to spatiotemporal knowledge bases

is added to each newspaper article. Using this information, the newspaper articles can be placed on maps or timelines. Through this spatiotemporal visualization, it is possible to grasp features or circumstances of an area by looking at the distributions and transitions of topics. This is the

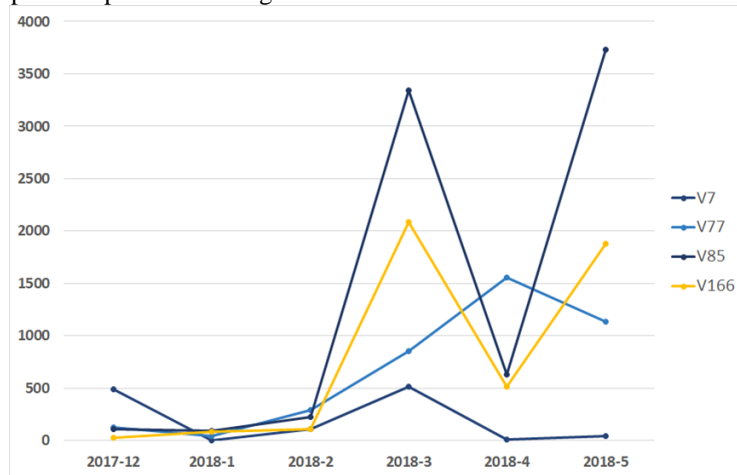


Figure 11: Transition of topics about 愛知県(Aichi Prefecture), 松山市(Matsuyama City)” and 今治市 (Imabari City)

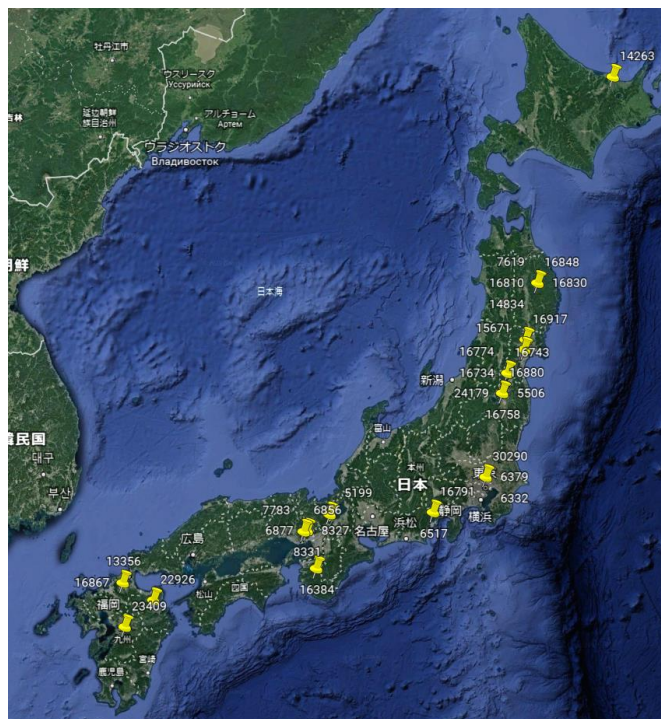


Figure 12: Distribution of the topic related to “Earthquakes” (Base Map: Google, 2019)

case study of a tool related to the title of this paper – “Prototyping Information System to Extract Area Study Information from Web Big Data.” Figure 12 is an example visualization of the distribution of news articles mostly related to topic 3 (V3) in the

first half of 2018, which includes words about earthquakes such as 地震 (earthquake), 津波 (tsunami), 避難 (evacuation), 復興 (recoveries), etc. This figure shows that these news articles are



distributed in Tohoku and Kyushu area that were heavily damaged by big earthquakes. Figure 13 shows the difference of distribution about news articles mostly related to topic 117 (V117) in December 2017 (left map) and May 2018 (right map), which includes words about disasters such as 地震 (earthquake), 雨 (rain), 雪 (snow), 台風 (typhoon), 噴火 (eruption), etc. The left map

contains news articles about snow and strong winds in northern Japan, Japan Sea coast, and mountain areas. The right map includes news articles about rainy seasons especially in Southern Japan. In these ways, visualization enables researchers to grasp overall views and trends in an area through the verification of specific topics.

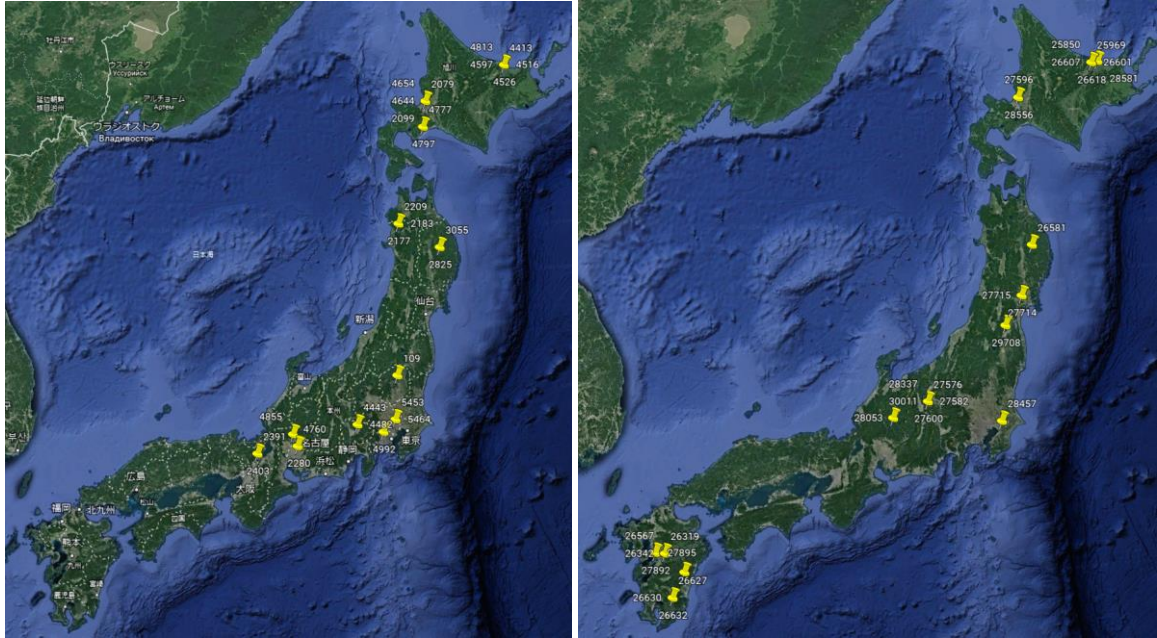


Figure 13: Difference of topic distribution related to “Disasters” in winter (left) and spring (right) (Base Map: Google, 2019)

### 3. Discussions

In area studies, there are some studies to collect and analyze newspaper articles on the Web (e.g., Yamamoto, 2012). There are also examples of social studies research, where, by using Big Data from Social Networking Service, the behavior patterns of residents are studied to draw up an optimal evacuation plan at times of disaster (e.g., Ming). In these studies, the target information sources are limited and collected based on skillfully pre-determined keywords.

However, when the objective is to provide a comprehensive overview of trends of an area and detect bursty structure, it is impossible to set keywords in advance, and big data on the Web should be input to analysis without being arbitrarily cut. “The development of an information system to automatically extract and visualize information that provides study hints from Big data on the Web,” the issue covered in this paper, exactly corresponds to this subject. In order to develop such an information system, this study selected newspaper articles on the Web as main source because these represent an

efficient, adequate and relevant source for collecting and analyzing data. For newspaper articles on the Web, this study tried to combine the following information processing technologies: (1) Web Scraping to collect data from the Web, (2) neural networks based on BiLSTM-CRF to segment words and recognize spatiotemporal expressions, (3) topic model based on the LDA to extract topics, and (4) Web application to visualize topics. At the point this paper was prepared, all modules described above have been processed separately. In the future, these processes will be integrated swiftly.

The data collection through Web Scraping for data used in this paper has been operated without any major problems. However, some other newspaper websites are behind paywalls, and it will be necessary to deal with this issue in order to build a more comprehensive database.

For word extraction and spatiotemporal expression recognition, the recognition accuracy has yet to be measured quantitatively. However, based on the learning curve (Figure 5) obtained from the experiment using training data, and the fact that the

structure of the training data and newspaper articles to be tested differed not so much, it is possible to consider this process functions enough for the next process. With manual error analysis, the following issues have been raised in regard to its practical use.

- (1) If the place name does not appear in an article explicitly and some named entities denote places implicitly, it is necessary to estimate its longitude and latitude using other spatial descriptions in the article.
- (2) If the place name extracted has corresponding multiple entries in knowledge databases, it is necessary to disambiguate using the context.

To give an example of those problems using Figure 2, if there is no entry for “諫早(Isahaya)” in our knowledge base, it is preferable to be able to estimate its latitude and longitude from surrounding distinctive phrases such as “堤防開門(Opening the gate of the bank),” which seems a general term but is actually used only for mentioning the Isahaya issue these days (Problelem (1) mentioned above). Additionally, for the latter example, as the place name “熊本(Kumamoto)” has multiple entries such as in “熊本県(Kumamoto Prefecture),” in “嘉麻市(Kama City),” in “北九州市(Kita-Kyushu City)” and in “田川郡(Tagawa County),” it is also preferable to determine the correct place intended (Problelem (2) mentioned above).

Moreover, extracted topics using the topic model has been confirmed to make sense and useful for grasping overall views and trends in an area. In this study, the LDA topic model was constructed by training newspaper articles of the fixed period and that model was used to detect topics of newly collected newspaper articles. This allowed achieving consistency over different time periods. However, as the time difference increases between the training data and the new data, the content may change even for similar topics (section 2.3.4), thus regular reconstruction of the topic model can improve the performance. There are some methods proposed as a dynamic topic model (e.g. Wang et al. 2008) and we are investigating the way to make visualized topics understandable and accurate at the same time. Additionally, this experiment was based on the Japanese language, but since the purpose of this study is to develop a system for area studies, the support for multiple languages is also essential. Because LDA is highly independent from language features, so that words sharing morphological roots can be automatically categorized in the same topic

(Schofield et al., 2017), it is not difficult to extend our implementation to multilingual application. As each newspaper is written in each language, LDA cannot detect correspondence between different languages. The use of a thesaurus would allow the problem to be handled.

In terms of visualization, at this moment, the information can be simply visualized, but this can be further developed by combining this first visualization attempt with other GIS tools, such as investigating the spatial relationships between topics by overlaying multiple topics as layers. Additionally, as topic model allocates multiple topics in one newspaper article, it is possible to develop a tool specific for topic model visualization, such as expressing the composition of the topics with different colors (Takada et al., 2014). Since the purpose of this study is to develop an information system applied to area studies, in addition to expanding the target languages from Japanese to English and Spanish, etc., the study plan includes the future inclusion of social media sites as data sources.

The future progress of this study must also take into consideration questions of what kind of visualization and method of analysis supports the awareness of area study researchers. Some analysis functions such as detecting changes in topic distributions and trends over space and time might be useful for new research hints.

### Acknowledgments

This study received support from the JSPS grants-in-aid for scientific research JP16H01897, Kyoto University Research Coordination Alliance Research Unit, and the Center for Southeast Asian Studies Kyoto University Glocal Information Network.

### References

- Ariya Hidayat: PhantomJS, <https://github.com/ariya/phantomjs>
- Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, 993-1022.
- CD-毎日新聞 2010~2015 データ集: <http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html>
- Graves, A., Mohamed, A. R. and Hinton, G., 2013, Speech Recognition with Deep Recurrent Neural Networks, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*. DOI: 10.1109/ICASSP.2013.663-8947.
- Griffiths, T. L. and Steyvers, M., 2004, Finding

- Scientific Topics, *Proc. of the National Academy of Sciences of the United States of America*, Vol.101, 5228-5235.
- Hara, S., 2010, Area Informatics – Concept and Status –, In: Ishida T. (eds) Culture and Computing. *Lecture Notes in Computer Science*, Vol. 6259, Springer, DOI [https://doi.org/10.1007/978-3-642-17184-0\\_17](https://doi.org/10.1007/978-3-642-17184-0_17).
- Hara, S., 2017, Digital Gazetteer as a Knowledgebase for Open Data Science, DOI: 10.23919/PNC.2017.8203524, 2017.The SQLite project: SQLite Home Page, <https://www.sqlite.org/index.html>
- Huang, Z., Xu, W. and Yu, K., 2015, Bidirectional LSTM-CRF Models for Sequence Tagging, arXivpreprint arXiv:1508.01991.
- HuTime: <http://www.hutime.jp/>
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. and Den, Y., 2014, Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, Vol. 48(2), 345-371.
- Ming-Hsiang, T., Spatiotemporal Modeling of Human Dynamics Across Social Media and Social Networks, <http://socialmedia.sdsu.edu/>
- Neubig, G., Nakata, Y. and Mori, S., 2011, 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.
- Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. 2017, Understanding text pre-processing for latent Dirichlet allocation, *ACL Workshop for Women in NLP (WiNLP)*.
- Wang, C., Blei, D.M., & Heckerman, D. 2008, Continuous Time Dynamic Topic Models, *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*.
- Yamada, T., 2017, A Study on Application of Topic Model to Newspaper Articles and Time Series Change of Topic. *IPSJ SIG Technical Report*, Vol. 2017-CH-115, No.1, 1-5 (Japanese).
- Yamamoto, H., 2012, Disaster Area Informatics Mapping System and Its Application, *IPSJ SIG Technical Report*, Vol.2012-CH-95, No.9,1-4 (Japanese).
- Takada, Y., Watanabe, H., Yanagisawa, M. and Yamada, T., 2014, A Visualization Method of Field Notes Based on Locations and Topic Models. *Proceeding of The Computer and the Humanities Symposium*, Vol. 2014, No.3, 57-62, (Japanese).