Development of Methods to Extract Place Names and Estimate Their Places from Web Newspaper Articles

Shoichiro Hara
Center for Southeast Asian Studies
Kyoto University
Kyoto, Japan
https://orcid.org/0000-0001-9343-1800

Akira Kubo
Graduate School of Arts and Sciences
The Open University of Japan
Chiba, Japan
https://orcid.org/0000-0002-4901-8290

Masato Matsuzaki
Faculty of Engineering
Kyoto University
Kyoto, Japan
matsuzaki.masato.86x@st.kyoto-u.ac.jp

Hirotaka Kameko
Academic Center for Computing and
Media Studies
Kyoto University
Kyoto, Japan
https://orcid.org/0000-0001-9844-6198

Shinsuke Mori
Academic Center for Computing and
Media Studies
Kyoto University
Kyoto, Japan
https://orcid.org/0000-0001-8596-8667

Abstract— **Area studies is an interdisciplinary study of the humanities adjacent to various research fields such as social sciences, natural sciences, engineering, medicine, and health. If focusing on humanity sciences, researchers' primary research resources have been text media, such as historical records, literary works, research papers, newspapers, and magazines. Researchers have engaged in analyses by reading the texts carefully as part of their research activities. However, with the spread of the Internet, Web data have become an inevitable source for area studies. We have explored new directions for area studies based on informatics compatible with big data and the Internet age. However, the low accuracy of place name extraction from texts and their place identification on a map hinders text processing, making it impossible to analyze big text data on the Web automatically. Using BiLSTM-CRF and a Balanced Corpus of Contemporary Written Japanese, our previous study realized approximately 0.9 of BiLSTM-CRF recognition accuracy, which demonstrated the effectiveness of BiLSTM-CRF. Our method can extract place names properly, but not enough if it is difficult to correctly identify their places on the map. Currently, we are trying a simple strategy: We assume that the locations of place names appearing in the same news article are close to each other. Thus, we compute the weighted average of all candidate coordinates of all place names in a news article as the pseudo center, then, as for each place name, choose the coordinate with the shortest distance from the pseudo center. This paper introduces the above methods and results in detail.**

Keywords— area studies, big data, machine learning, Named Entity Recognition, Geocoding, LDA, KyTea, BiLSTM-CRF

I. INTRODUCTION

Area studies is an interdisciplinary study of the humanities adjacent to various research fields such as social sciences, natural sciences, engineering, medicine, and health. If focusing on humanity sciences, researchers' primary research resources have been text media, such as historical records, literary works, research papers, newspapers, and magazines. Researchers have engaged in analyses by reading the texts carefully as part of their research activities. With the spread of the Internet, various types and large volumes of information related to area studies are now being distributed on the Web. This trend has become even more prominent in large-scale disasters, political conflicts or changes, and national elections. Even if only limited to text media analysis, there is already a massive volume of data on the Web. It has become impossible for humans to read and analyze all the extracted data. This paper explores new directions for informatics-based area studies compatible with big data and the Internet age to address these issues.

Our study has used Internet news articles to extract information related to area studies. The reasons for using newspapers are that, (1) comparing to social media, such as Twitter or Facebook, the texts in newspapers are more consistent in terms of the use of grammar and vocabulary, which makes it easier to apply natural language processing; (2) newspapers are comprehensive information sources for collecting information on many topics such as politics, economics, and culture; and (3) most of the newspaper's companies publish news articles not only in paper media but also as digital data on the Web, which allows for efficient data collection and processing. Some research in area studies have used Web text data[1,2], but their purposes are limited to specific interests; that is, researchers can retrieve necessary news articles for their analyses using skillfully pre-determined keywords. This method is helpful for researches with clear and narrow purposes, such as political changes and regional disasters. However, when the purpose of area studies is to reconstruct a comprehensive view of an area, such as "What kind of area is Taiwan or what will happen to Taiwan in the future?", it is impossible to prepare for the right keywords in advance. Therefore, our previous study [3] developed an information tool that automatically categorizes news articles using Topic Models based on the LDA (Latent Dirichlet Allocation [4]). Experiments and evaluations using Mainichi Shimbun newspaper articles (CD−Mainichi Shimbun, published during six years from 2010 to 2015, number of newspaper articles: 606,924, number of different terms: 2,683,289, the total number of words: 286,288,248) showed this method's effectiveness in categorizing news articles automatically.

However, the low accuracy of place name extraction from news articles and their place identification on maps hinders the automatic text processing by our tools. The former problem is related to the choice of the right word for a geographical context. For example, a word "九条 (Kujo or Kyujo)" appears not only as a geographical reference (e.g., 九条通: Kujo Street) but also in different contexts such as a family name (e.g., 九条家:Kujo Family), a crop name (e.g., 九条葱:Kujo leek), and as a legislative article (e.g., 憲法九条:Article 9 of the Constitution). The latter is the problem of selecting the correct place from different places with the same name. For example, there are many places named "三条 (Sanjo)" in Japan, such as "三条市(Sanjo City)" in Niigata prefecture (37.636778, 138.961667), "三条通(Sanjo Street)" in Kyoto City (35.0084, 135.7471 around the midtown of Kyoto City) and "三条(Sanjo: old town name)" in Nagoya City (35.107578, 136.902958).

In order to solve the problems cited above, i.e., to extract place names from texts correctly, we have used BiLSTM-CRF[5], combining a Bi-directional Long Short Term Memory (BiLSTM) network and Conditional Random Fields (CRF). BiLSTM-CRF is known for achieving high precision in recognition of named entity tasks. Our previous study used a Balanced Corpus of Contemporary Written Japanese (BCCWJ) in which we manually added some tags to indicate place names as a learning data set to train BiLSTM-CRF. In this estimation experiment, the BiLSTM-CRF recognition accuracy was approximately 0.9, which demonstrated the effectiveness of BiLSTM-CRF. Though our method can extract place names appropriately from texts, it is not enough to identify the correct place. Currently, we are trying a simple strategy: assume that the locations of place names appearing in the same news article are close. Thus, we compute the weighted average of all candidate coordinates of all place names in a news article (hereafter, pseudo center). If a place name has some candidate coordinates, from the assumption, this coordinate should be close to another place name, i.e., close to the pseudo center. Thus, we choose the coordinate with the shortest distance from the pseudo center for each place name.

In this paper, Chapter II will describe methods of preparing for the news article data, training the LDA topic model, extracting place names from news articles, and identifying their locations on maps. Chapter III will show the results, and lastly, the problems and final considerations will be discussed in Chapter IV.

II. Methods

A. Data Collection

As in the previous study, this study also collected news articles from Mainichi Shimbun (Web aggregation and commercially available CD-ROM), Asahi Shimbun (Web aggregation), Yomiuri Shimbun (Web aggregation), and AFP (English and Spanish news articles via AFP Forum), and organizes the following information:

- ID
- Article title
- Article main text
- Published date/time
- Article URL (URL for the Web page which includes the main text of the newspaper article)

News articles aggregated from the Web are cleansed, broken each HTML text into components, and reformatted as the CSV data following the above structure. The character encoding for this text file is UTF-8 (without BOM). Since Mainichi Shinbun news articles in CD-ROM have a key-value structure, a simple mapping operation can convert this data to the SCV data.

As a preliminary study, this study used news articles from Mainichi Shimbun CD-ROM Editions, officially provided by The Mainichi Newspapers Co., Ltd (national edition and 46 local editions; published ten years from 2010-01-01 to 2019-12-31) [6].

B. LDA Topic Model

As a method to automatically infer subjects of news articles, this study uses the LDA topic model. The LDA topic model expresses a topic as the sets of words that are statistically likely to co-occur.

The LDA topic model needs Bag-of-Words (BOW) for training. In BOW, a news article is represented as the set of its words, disregarding grammar and even word order but keeping multiplicity. Therefore, we segmented news articles into words using the KyTea[7], a Japanese morphological analyzer, and secondly, chose nouns, verbs, adjectives, adjectival nouns, and adverbs that characterize news articles. We selected the most frequent 100,000 words used only in less than 50% of news articles. The statistical of the extracted corpus is shown in TABLE Ⅰ.

TABLE I. Corpus specifications of the Mainichi Shimbun CD-ROM

	Articles	Sentences	Words	Characters
National	607,671	7,791,338	219,677,468	321,075,103
Local	1,338,053	14,032,449	401,913,172	587,034,329
Total	1,945,724	21,823,787	621,590,640	908,109,432

We used the gemsim [8], a python package of the LDA, for implementation. We set the parameter α as

$$\alpha_i = \frac{1.0}{i + \sqrt{N}} ,$$

where N is the number of topics, and *i* shows the index of the topic. We used the default parameters for all other hyper-parameters.

The LDA topic model uses two parameters, Perplexity and Coherence, to determine the appropriate number of topics. Perplexity represents the number of branches or choices. A smaller value indicates a better model. Coherence is an index to measure the model's generalization performance, and a larger value is considered to be indicative of a better learning algorithm. We trained the LDA topic model by changing topic numbers from 16 to 256 in 8 increments and calculated Perplexity and Coherence. At the same time, we interpreted frequent vocabularies constituting each topic, then finally determined the topic number as 184.

C. Named Entity Recognizer

In order to recognize spatial expressions in texts, we designed a named entity recognizer (NER) for them with our entity definition.

a) Training Method: We adopted the BiLSTM-CRF for the task of NER task. We used 5-fold cross-validation for evaluation because the amount of training data is limited. In each sub-sample, we divide news articles into train:dev:test =

3:1:1, train model on training data, and then choose the epoch achieving the best F-measure for development data. A model with the best F-measure over all sub-samples is used for the final model as input for the longitude-latitude estimator.

b) Training Data: We selected 100 news articles each about earthquakes and floods and annotated them. We listed the target news articles by choosing news articles with the highest probabilities for the chosen topic. Note that we selected news articles over 4,096 bytes.

c) Annotation Standard: We categorize spatial expressions into two types: absolute expression and relative expression. Previous entity tagging schemes regarding location [9,10] focus on absolute expression, and relative expression is overlooked. Relative expressions are valuable for inferring the author's intent regarding location; therefore, we establish an annotation standard for both absolute and relative spatial expressions. We establish five absolute expression tags.

L: Locations and facilities (e.g., Olympic Games were held in Tokyo).

LD: Locations in disaster names (e.g., Great East Japan Earthquake).

LF: Locations included in a word that is not a spatial expression (e.g., New Yorkers).

LO: Locations meaning organizations but impossible to determine by context.

LV: It may indicate a location, but the context is ambiguous and cannot be determined. Except for LO.

We also establish five relative expression tags.

RLI: Inside of location (e.g., northern Kyoto).

RLO: Outside of location (e.g., about 450km northwest of Okinawa Island).

RLA: Neighborhood of location (e.g., around Kyoto).

RLB: Between locations (e.g., between Tokyo and Osaka).

RL*: Expressions other than those above.

D. Longitude-Latitude Estimator

The longitude-latitude estimator is a module to estimate an absolute place for each NE (named entity). In this study, the absolute value is a pair of longitude and latitude, which allows us to point the NE to a single point on the globe's surface.

The estimation process is composed of two stages. The first step processes the division of compound location entities into individual expressions. For example, "神奈川・千葉両県 (both Kanagawa and Chiba prefecture)" is divided into "神奈川県 (Kanagawa prefecture)" and "千葉県 (Chiba prefecture)." Conversion rules generated by observation from the news articles do this. Then, we search geo-coordinates in the gazetteer database by entity names. The gazetteer database contains locations and positions collected from the Historical Gazetteer Data[11] and Wikidata[12]. It also contains the annotated geolocation data we used to expand the learning data set of BiLSTM-CRF, where we added the correct position (longitude and latitude) to the tags used to indicate a place name. It is geocoded using Wikidata, OpenStreetMap[13], and Community Geocoder[14].

The search considers both the exact match and the partial match. If we find multiple candidates for an entity, we score them and choose the best one. I.e., we compute the average of candidate coordinates for each entity and then the average of all entities (pseudo center). Then, we measure the distance between the pseudo center and candidate coordinates for each entity and add a penalty if a candidate is found by partial-match. Finally, we choose the candidate with the shortest distance.

In estimation, we give the output from NER for an entire news article to the estimator. The pseudo center is computed by using all candidate coordinates in the news article and used globally.

III. RESULTS

A. LDA Topic Model

This section will try to interpret the results of the LDA topic model, focusing on the first one-year transition of topics related to the 東日本大震災 (2011.3.11 Great East Japan Earthquake. Hereafter the Earthquake). As the topics seemed to be related to the Earthquake, we chose three topics; 37, 110, and 124. The frequent top five words assigned to each topic are as follows:

- **Topic 37**: 災害/名詞, 避難/名詞, 防災/名詞, 地震/名詞, 被害/名詞 (disaster/noun, evacuation/noun, disaster prevention/noun, earthquake/noun, damage/noun)
- **Topic 110**: 原発/名詞, 電力/名詞, 発電/名詞, 原子/名詞, 放射/名詞 (nuclear power/noun, electric power/noun, power generation/noun, atom/noun, radiation/noun)
- **Topic 124**: 震災/名詞, 被災/名詞, 地/名詞, 日本/名詞, 復興/名詞 (earthquake/Noun, damage/Noun, place/Noun, Japan/Noun, reconstruction/Noun)

We interpreted Topic 37 as the occurrence of disasters, associated damages, evacuations, and the like. Likewise, Topic 110 as nuclear accidents, associated power losses, radiation leaks, and the like. Topic 124 is similar to Topic 37, but we interpreted it as having implications for reconstruction and assistance. Fig.1 shows how these three topics' composition changed in a year after the Earthquake across several regions. Each topic value results from the Moving-Average calculation whose time-period is seven days; each topic value is obtained by first accumulating the topic ratio of all news articles published during the time-period and secondarily dividing the accumulation by the number of news articles.

We can see from this figure that the damage situations in the disaster areas and their surrounding areas were different, or even in the disaster areas, the situations were also quite different between the places where the tsunami damages were dominant and the places where nuclear accidents were dominant.

In the case of the national edition of Mainichi Shimbun, the ratios of all three topics rose sharply immediately after the Earthquake and then gradually declined. Not shown in this figure, these topic ratios continued to decrease except a slight increase every March, the month of reference for when the Earthquake occurred. After ten years, it returned to the same level before the Earthquake.

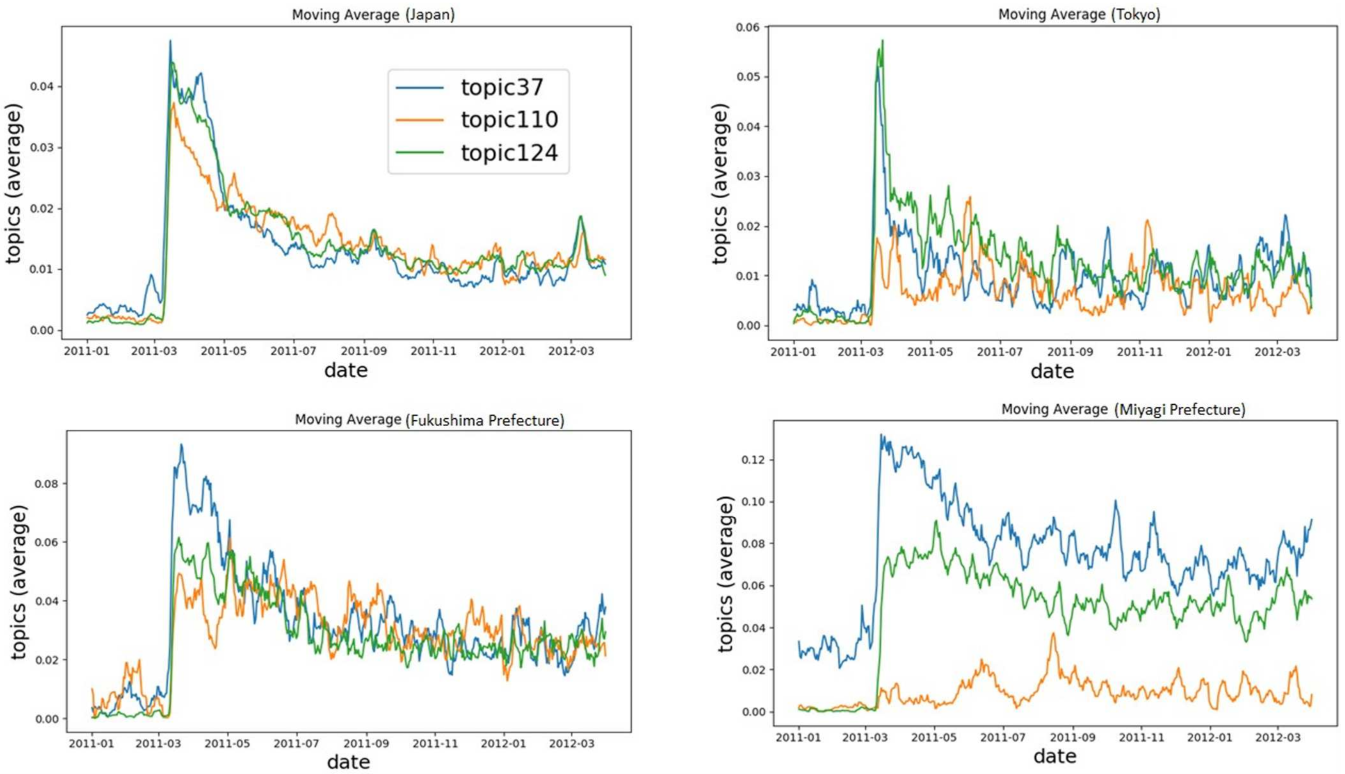


Fig. 1 Change of three topics in one year across different areas. (upper left) news articles from the national editions. (upper right) regional editions of Tokyo. (lower left) regional editions of Fukushima prefecture, which was suffered from the atomic reactor accidents. (lower right) regional editions of Miyagi prefecture, which is one of the most damaged areas by the Earthquake and Tsunami. (Blue-lines) topic 37; disaster, evacuation, disaster prevention, earthquake, damage, etc. (Orange lines) topic 110; nuclear power, electric power, power generation, atom, radiation. (Green lines) topic 124; earthquake, damage, place, Japan, reconstruction etc.

The ratios of the three topics in the Tokyo edition are similar to those in the national edition. However, the ratio of the Topic 110 on nuclear reactors and radiation is slightly superior. Tokyo area was also affected by the Earthquake, but to a lesser extent than in the Earthquake-stricken areas. However, the spread of radioactive contamination due to the nuclear accident became a big issue. Actually, at some of the peaks seen in Topic 110, there were news articles about implementing independent radiation dose measurements by local governments, rapid increases in dose in some cities, or finding radiation hotspots.

The decreasing trend of the three topic ratios in the Miyagi and Fukushima editions was slight, unlike the national and Tokyo editions. The reason may be that the Earthquake severely damaged both areas more than Tokyo.

TABLE Ⅱ. Result of NER performance on test data. Answer and Result are the number of occurrences on test data and predicted data, respectively. True is the number of successful predictions. Result of tag LV, LF, and RL* is not shown because they have not occurred in test data.

Tag	Answer	Result	TRUE	Precision	Recall	F-measure
L	2521	2253	1949	86.51	77.31	81.65
LD	476	480	454	94.58	95.38	94.98
LO	57	10	1	10.00	1.75	2.99
RLI	274	244	200	81.97	72.99	77.22
RLA	73	56	40	71.43	54.79	62.02
RLB	22	26	18	69.23	81.82	75.00
RLO	19	2	1	50.00	5.26	9.52
ALL	3442	3071	2663	86.71	77.37	81.77

On the other hand, the trends in topic ratios in both prefectures are pretty different, which may reflect that the tsunami heavily damaged Miyagi prefecture, and the number of fatalities and missing persons was higher than that of Fukushima prefecture. At the same time, Fukushima prefecture was greatly affected by the nuclear accident. Therefore, there are many news articles about earthquake damages and reconstructions in the Miyagi edition, and this tendency has continued for several years. However, there are a few news articles about nuclear accidents and radiation. On the other hand, in the Fukushima edition, Topic 110 related to the nuclear accidents was outstanding, and the number of related news articles increased two years after the Earthquake.

B. Named Entity Recognizer

As parameters, we used: a word embedding size of 300 and used all occurrence words for embedding; a hidden layer size of 256 with a batch size of 30; and a learning rate of 1e-3 with a weight decay of 1e-6 for the SGD (stochastic gradient descent) optimizer. We trained the model up to 100 epochs on each sub-samples and choose the best epoch using the F-measure of development data and use it for prediction. We evaluate precision, recall, and the F-measure by using test data of all sub-samples.

The performance result is shown in TABLE II, and the confusion matrix of tags is shown in Fig.2. It is the best F-measure in tag sets for the tag LD because disaster names are

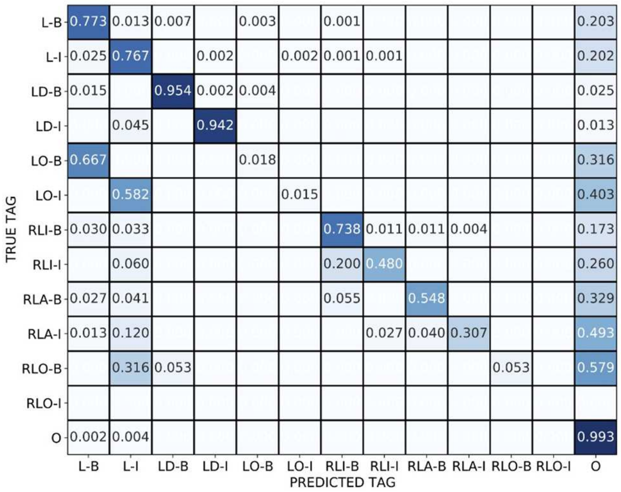


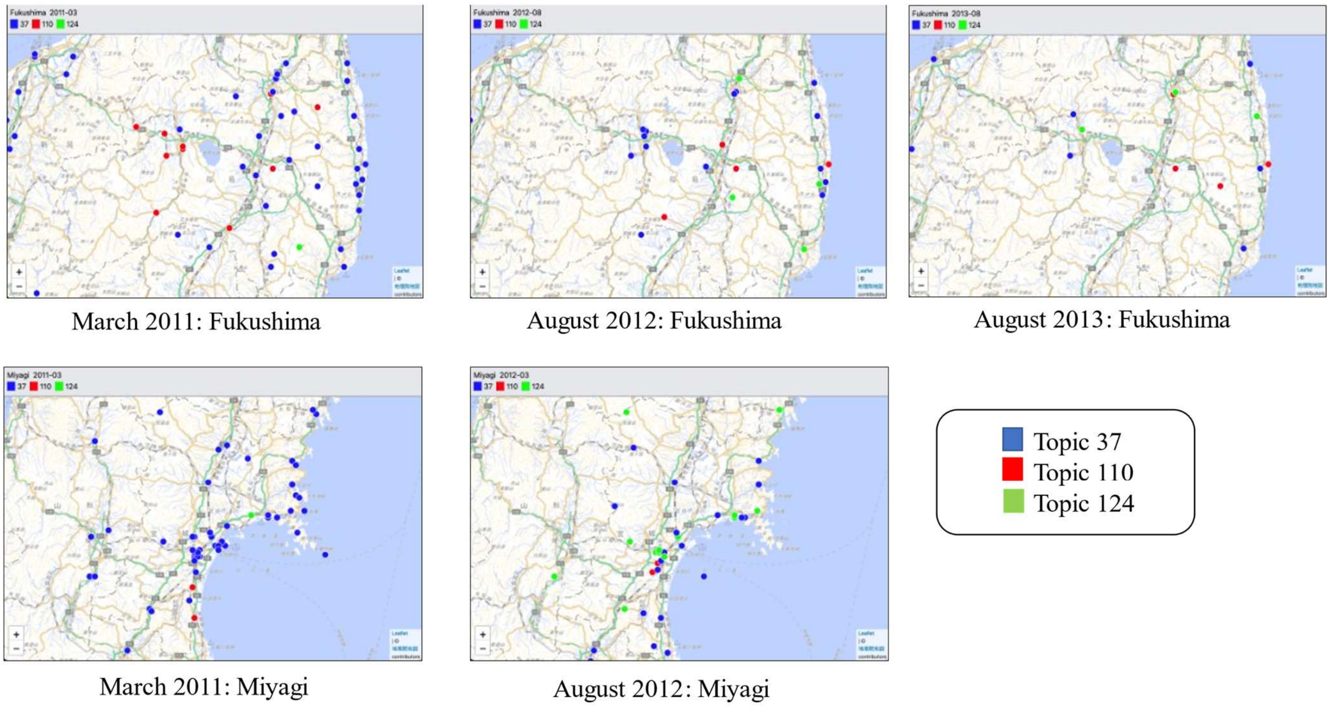
Fig.2. Confusion matrix of tags. The confusion matrix shows the relationship between the true tag and the predicted tag that is the result of recognition. Each row represents prediction occurrences for the true tags. In an ideal situation, only the diagonal component of a matrix would be non-zero, otherwise zero. The more non-zero there are other than diagonals, the more likely the tag is to be mispredicted. In addition, the confusion matrix shows how the recognizer tends to tag incorrectly. For example, there is a large value (filled with darker blue) in the off-diagonal L tag column for the LO tag, indicating that the LO tag tends to be incorrectly predicted as the L tag.

determined by the government and local authorities and have expressions with less variation. In contrast, the F-measure of the LO and the RLO tag is very low. For the LO tag, administrative locations are predicted as L tag in many cases. For example, "今回/O の/O 姫路/LO-B 市/LO-I の/O 取り組み/O で/O は/O ... (Himeji City's latest project is ...)" is tagged as "今回/O の/O 姫路/L-B 市/L-I の/O 取り組み/O で/O は/O ...." For the RLO tag, the RLO tag following the L tag tends to be predicted as the L tag. For example, "ＪＲ/L-B 豊岡/L-I 駅/L-I 前/RLO-B (in front of JR Toyooka Station)" is tagged as "ＪＲ/L-B 豊岡/L-I 駅/L-I 前/L-I."

C. Longitude-Latitude Estimator

We randomly choose 50 news articles from 873 news articles that include topic numbers 37, 110, or 124 in the Miyagi and Fukushima editions. The news articles are not used for NER training and evaluation. We computed the accuracy of coordinate estimation for successfully tagged entities in the NER stage, and the result was 71.43%. Successfully estimated coordinates come from Wikidata and the annotated geolocation data. In some cases, different coordinates are estimated for the same location. For example, "福島" (Fukushima) does not refer only to one place because the name is also used in other places outside Fukushima prefecture, such as Osaka. Such coordinates can affect the pseudo center and result in instability and performance drop.

Fig.3 shows the geographical distribution of news articles estimated by the above method. The upper row shows the topic distribution of Fukushima prefecture, which was affected by the nuclear accident, and the lower lowes shows that of Miyagi prefecture, which suffered tsunami damages. The left column shows the distribution in March 2011, just after the Earthquake happened, the middle column shows the distribution in August 2012, almost one year and a half after the Earthquake.



March 2011: Fukushima

August 2012: Fukushima

August 2013: Fukushima

March 2011: Miyagi

August 2012: Miyagi

Fig. 3 Spatial distribution changes of three topics in Fukushima and Miyagi prefecture. (upper left) Fukushima prefecture in March 2011, just after the Earthquake, (upper middle) Fukushima prefecture in March 2012, one year after the Earthquake. (upper right) Fukushima prefecture in August 2013. (lower left) Miyagi prefecture in March 2011, just after the Earthquake, (lower middle) Miyagi prefecture in March 2012, one year after the Earthquake. (Blue dots) topic 37; disaster, evacuation, disaster prevention, earthquake, damage, etc. (Red dots) topic 110; nuclear power, electric power, power generation, atom, radiation. (Green dots) topic 124; earthquake, damage, place, Japan, reconstruction etc.

In Fukushima prefecture, Topic 37 (related to disasters) points were decreased and distributed in specific areas; on the other hand, Topic 124 (related to reconstruction) points increased. In Miyagi prefecture, although the distribution of Topic 37 points increased like Fukushima prefecture, Topic 24 points were still distributed to many places, mainly in coastal areas. Above right is the topic distribution of Fukushima in August 2013, more than two years after the Earthquake. The upper right is the topic distribution of Fukushima in August 2013, more than two years after the Earthquake. As mentioned in III a), the ratio of Topic 110 increased from this time. Although detailed investigation will be needed, it is presumed that the Miyagi edition began publishing radiation dose measurements of some specific areas regularly, which increased the number of news articles related to Topic 110.

IV. CONCLUDING CONSIDERATIONS

The original motivation of this research was to look for new methods of area studies within the domain of informatics by using advanced information methods. Therefore, we addressed the issues of (1) targeting a large amount of data that researchers cannot read and analyze manually and (2) establishing a method for comprehensively grasping an area beyond a specific viewpoint. As the first step, we have been developing a method to classify a large amount of text data based on topic (or subject), geographical position, and time. Regarding the classification of text by topics, our previous study verified the effectiveness of LDA.

This study tried to visualize the distributions of topics on a map. We demonstrated that BiLSTM-CRF could predict named entities regarding location with approximately 0.8 F-measure for named entity recognition. We also reported the trend of error prediction and tag ambiguity. For longitude-latitude estimation, we demonstrated that a simple strategy that uses a gazetteer database could achieve approximately 70% accuracy and explored the limitations of this strategy. We expect that future work will improve tag disambiguation, make the coordinates selection method more robust, and explore the usefulness of relative location expression to improve geocoding and visualization.

We have promoted the modularization of methods for collecting news articles from the Web, cleansing messy text data, recognizing and extracting tokens including named entities for further analysis, topic analysis, and visualizing topics on a map, which have been realized in the research so far. We are building an easy-to-use analysis tool by putting these modules together.

Time as much as location is essential information for area studies but has not been covered in our studies. Our research sources have been news articles, and we could use the publication date (though this premise is not always applicable to all news articles). However, as for general text data, it is necessary to extract time-related words and convert them to a standard format such as ISO 8601. We are considering the possibility of applying BiLSTM-CRF to predict time-related words. In addition, a tool equivalent to gazetteer databases that organize place names and their latitudes and longitudes has already been developed. Supposing that we could process time information using these methods, in that case, we believe that our first object, which is "developing a method to classify a large amount of text data based on topic, geographical position, and time" will be achieved.

ACKNOWLEDGMENT

This study received support from the JSPS KAKENHI Grant Number JP21H04376, Kyoto University Research Coordination Alliance Research Unit, and the Center for Southeast Asian Studies Kyoto University Glocal Information Network.

REFERENCES

[1] Wacana Informasi Bencana Alam dan Keadaan Sosial, http://disaster.net.cias.kyoto-u.ac.jp/Indonesia/ (accessed on 2021/Aug/14)

[2] Spatiotemporal Modeling of Human Dynamics Across Social Media a nd Social Networks, http://mappingideas.sdsu.edu/election/ (accessed on 2021/Aug/14)

[3] S. Hara, T. Yamada, M. Ishikawa, K. Shirai, A. Kameda, and S. Mori, "Prototyping Information System to Extract Area Study Information f rom Web Big Data," International Journal of Geoinformatics, Vol. 15, Issue 2, pp.57-67, Apr-Jun2019.

[4] D.M.Blei, A.Y.Ng, and M.I.Jordan, "Latent Dirichlet Allocation," Jou rnal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[5] Huang, Zhiheng, Wei Xu, and Kai Yu, "Bidirectional LSTM-CRF Mo dels for Sequence Tagging," arXivpreprint arXiv:1508.01991, 2015.

[6] The Mainichi Newspapers Co., Ltd., https://www.mainichi.co.jp/com pany/corporateprofile-e.htm (accessed on 2021/Aug/14)

[7] Graham Neubig, Yosuke Nakata, and Shinsuke Mori, "Pointwise pred iction for robust, adaptable Japanese morphological analysis," In Proc eedings of the 49th Annual Meeting of the Association for Computati onal Linguistics: Human Language Technologies, pages 529–533, Por tland, Oregon, USA, June 2011. Association for Computational Lingu istics

[8] Radim ˇReh˚uˇrek and Petr Sojka, "Software Framework for Topic M odelling with Large Corpora," In Proceedings of the LREC 2010 Wor kshop on New Challenges for NLP Frameworks, pages 45–50, Vallett a, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[9] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata, "Extended name d entity hierarchy," Proceedings of the Third International Conference on Language Resources and Evaluation (LREC '02), pages 1818–182 4, 2002.

[10] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, S ameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kau fman, Michelle Franchini, et al., "Ontonotes release 5.0 ldc2013t19," Linguistic Data Consortium, Philadelphia, PA, 23, 2013.

[11] Historical Location Data. https://www.nihu.jp/ja/publication/source_map (accessed on 2021/Aug/2)

[12] Wikidata. https://www.wikidata.org/(accessed on 2021/Aug/10)

[13] OpenStreetMap. https://www.openstreetmap.org/(accessed on 2021/A ug/10)

[14] Community Geocoder. https://community-geocoder.geolonia.com/(ac cessed on 2021/Aug/10)