Utilizing Geographic Entity Information for PLM-based Document Geolocation Models

Yuya Yamamoto, Takashi Inui

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Utilizing Geographic Entity Information for PLM-based Document Geolocation Models

Yuya Yamamoto [*]
College of Information Science
School of Informatics
University of Tsukuba
s2012003@u.tsukuba.ac.jp

Takashi Inui
Division of Information Engineering
Faculty of Engineering, Information and Systems
University of Tsukuba
inui@cs.tsukuba.ac.jp

Abstract

In the task of document geolocation, which involves estimating the posting location of SNS posts, mentions of place names (e.g., "Tokyo") or landmarks (e.g., "Disneyland") within the document often serve as strong clues. However, relying solely on these mentions does not always provide sufficient information. In this study, to utilize these mentions more effectively, we aim to identify the real-world entities that these mentions refer to and leverage the information associated with the identified entities. Through experiments, it was confirmed that incorporating entity information, specifically focusing on the location information of entities, into the document geolocation model improves the performance of document geolocation.

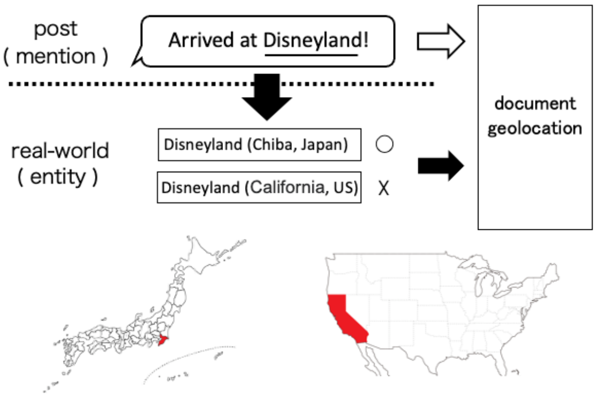

Figure 1: Utilizing real-world entity information, Disneyland(Chiba, Japan), in the document geolocation model. The white arrow is a regular input. The black arrows are additional inputs proposed in this work.

1 Introduction

Recently, social networking services (SNS) have become highly widespread, and SNS posts with location information are an essential source for social sensing. However, only a subset of SNS posts actually include location information, posing a significant challenge. To address this issue, research on document geolocation has been conducted, which aims to estimate the corresponding location information for SNS posts that do not have location information (Bo et al., 2012; Lau et al., 2017; Okajima and Iwakura, 2018a; Huang and Carley, 2019; Hasni and Faiz, 2021).

In document geolocation, mentions of place names or landmarks within the document often serve as strong clues. However, relying solely on these mentions does not always provide sufficient information. For example, suppose a traveler visiting Tokyo Disneyland in Chiba Prefecture posts on SNS, "Arrived at Disneyland!". While the posting location is expected to be related to the mention "Disneyland," for the mention "Disneyland" to serve as a compelling clue, it is desirable that the document geolocation model understands whether it refers to Tokyo Disneyland in Chiba Prefecture, Japan, or Disneyland in California, US.

Although there are studies that utilize external knowledge for document geolocation (Miyazaki et al., 2018; Hirakawa and Inui, 2022), discussions that focus on identifying real-world entities and leveraging their information have not been conducted. Therefore, this study focuses on identifying entities from mentions within the document and utilizing the information of the identified entities for document geolocation (Figure 1). Specifically, by adopting a pre-trained language model (PLM) for document geolocation, we will discuss which types of entity information should be incorporated into the model, how to convert this information into embedded representations, and how to integrate them into the model.

[*] Currently at GEN Co .,Ltd

2 Related Work

Document geolocation estimates the geographical location from which an input document, such as an SNS post, was posted. This task has been pursued since the 2010s, coinciding with the rise of SNS services, and has been actively discussed in Western languages, including being featured as a shared task in WNUT2016(Han et al., 2016), VarDial2020(Gaman et al., 2020), and VarDial2021(Chakravarthi et al., 2021).

Early document geolocation methods primarily focused on words within the input document, proposing techniques such as selecting words that are effective for classification(Bo et al., 2012) and filtering words(Morikuni et al., 2015). For Twitter data, studies have also utilized hashtags as features(Chi et al., 2016). With the proliferation of deep learning, various models and network architectures have been employed for this task, including methods using word embeddings(Miura et al., 2016), CNN-based methods(Fornaciari and Hovy, 2019), LSTM-based methods(Mahajan and Mansotra, 2021), and BERT-based methods(Scherrer and Ljubešić, 2021). Additionally, there has been research into incorporating supplementary information beneficial for classification into deep learning-based models in addition to the information from the input documents. The deepgeo model proposed by Lau et al.(Lau et al., 2017) is an LSTM-CNN-based neural model that utilizes not only the input SNS post but also the posting time and the location information from the user's profile.

While various studies have explored models and features effective for the document geolocation task, no research has thoroughly examined the effectiveness of entity information, as explored in this study.

3 Basic elements

Before delving into the main content of this paper, the components of this study will be explained.

Geographical Entities: In document geolocation, geographical entities related to locations, such as place names and landmarks, are considered particularly important. Therefore, this study focuses specifically on **geographical entities**. Specifically, among the entities included in the Japanese Wikipedia Entity Vectors published by Tohoku University[1], we use entities categorized as organization names, place names, facility names, and event names, according to the Extended Named Entity labels of the SHINRA Project[2][3] as geographical entities.

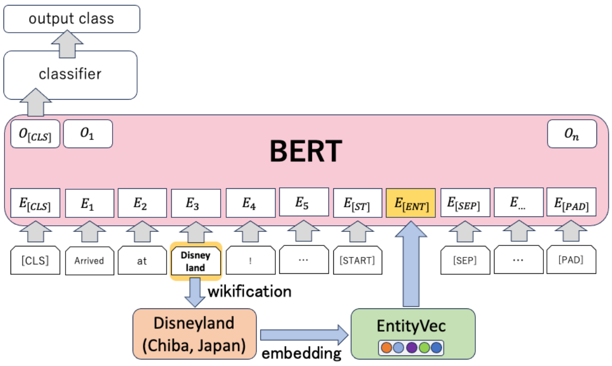

Figure 2: An example of incorporating entity information. The entity information (Disneyland(Chiba, Japan)) obtained through Wikification is converted into embedding representations and input as additional information into BERT.

Document Geolocation model: This study addresses the task of document geolocation at the Japanese prefecture level, where the goal is to output one of the 47 prefecture classes in Japan for the input document. For example, in the aforementioned case of "Arrived at Disneyland!", Chiba Prefecture would be the expected output class. For the document geolocation model, we adopt BertForSequenceClassification[4] available from Huggingface[5] as the base model, which is a document classification model based on BERT (Devlin et al., 2019). The detailed settings of this model are shown in Appendix A.1.

Entity Linking: The task of linking a mention within a document to a real-world entity is known as the entity linking task, with active research particularly in the area of Wikification, where Wikipedia pages are assumed as entities (Mihalcea and Csoma, 2007). This study also assumes Wikification and incorporates information from Wikipedia pages as entity information into the document geolocation model. The Wikipedia data used[6] was obtained from dump data in August

[1] https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

[2] https://2022.shinra-project.info/

[3] http://ene-project.info/

[4] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

[5] https://huggingface.co/

[6] https://dumps.wikimedia.org/other/cirrussearch/

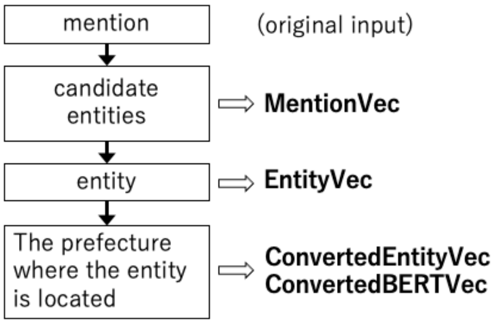

Figure 3: Embedding representation acquisition methods. ConvertedEntityVec and ConvertedBERTVec are novel methods proposed in this study, whereas EntityVec and MentionVec are approaches adopted from existing research.

2023.

4 Incorporation of Entity Information

An example of incorporating entity information into the document geolocation model is shown in Figure 2. This figure illustrates the case where entity information Disneyland(Chiba, Japan) is obtained from the mention "Disneyland" through Wikification.

In this study, when entities corresponding to mentions are given, we consider the following four methods for acquiring embedding representations from entity information. The differences in the embedding representation acquisition methods are summarized in Figure 3. Among these, **EntityVec** and **MentionVec** are methods adopted in existing research. On the other hand, **ConvertedEntityVec** and **ConvertedBERTVec** are novel methods proposed in this study specifically for the document geolocation task.

- **EntityVec**(Suzuki et al., 2016)

  In the case of **EntityVec**, embedding representations are acquired from the lemma of the entity(namely, the Wikipedia page). For implementation, Japanese Wikipedia entity vectors are used. These vectors are learned using word2vec(Mikolov et al., 2013), which takes into account the link information in Wikipedia.

- **ConvertedEntityVec**

  Our preliminary investigations confirmed that documents containing prefecture names can achieve good geolocation performance without incorporating entity information. Therefore, instead of using the lemma of the entity like in EntityVec, **ConvertedEntityVec** utilizes the entity information of the prefecture where the entity is located (prefecture entity). After converting the original entity into a prefecture entity, the process is the same as EntityVec's. The prefecture where the original entity is located is determined by referring to Okajima et al. (Okajima and Iwakura, 2018b), and is defined as the first prefecture mentioned in the main text of the Wikipedia page of the original entity.

- **ConvertedBERTVec**

  Similar to ConvertedEntityVec, **ConvertedBERTVec** utilizes the prefecture information of the entity's location. However, instead of using EntityVec to acquire embedding representations as in ConvertedEntityVec, the prefecture name is inserted into the original input text for BERT. This operation converts the inserted prefecture name, along with the original text, into embedding representations through the BERT training process.

- **MentionVec**(Kageyama and Inui, 2022)

  As a comparison method to verify the effectiveness of entity information, in the case of **MentionVec**, embedding representations are acquired from the information of the entity candidates rather than the identified entities. Specifically, embedding representations are obtained using EntityVec for each entity candidate, and the average vector of these embeddings is used.

As the incorporation positions for the embedding representations acquired through one of the above methods, we consider the following two types.

- **concat**

  A special token, "START," is added to the end of the token sequence, and the entity information is incorporated after this token. This method is based on Nakamoto et al. (Nakamoto et al., 2023). If there are multiple pieces of entity information, they are incorporated in the order of their appearance. Figure 2 provides an example of incorporating EntityVec using **concat** method.

- **infuse**

  For the token sequence, a special token "MENTION" is inserted just before the mention, and the entity information is incorporated between "START" and "END" immediately after the mention. This method is based on Faldu et al. (Faldu et al., 2021).

5 Experiments

5.1 Settings

5.1.1 Models

We construct models that incorporate entity information using the method described in the previous section, based on the BERT-based document classification model described in Section 3. We then compare the performance of these models.

5.1.2 Dataset

We used the Japanese Twitter posts dataset in the tourism domain (Hirakawa and Inui, 2020). This dataset consists of Japanese tweets posted from all 47 prefectures of Japan between 2014 and 2015. The prefecture information of the posting locations was used as the correct labels, obtained by reverse geocoding the geotags attached to all posts. The number of documents in the dataset is 197,741 for the training data, 4,000 for the validation data, and 7,000 for the evaluation data.

5.1.3 Mentions and Entities

The target mentions for entity information retrieval were defined as named entity classes representing locations, extracted by analyzing the documents using GiNZA[7] [8] .

Next, following the procedure from Kageyama et al. (Kageyama and Inui, 2022), for a given mention m, the set of entities $E(m)$ linked by m as anchor text within Wikipedia pages was used as the candidate entities for m. Entities that meet the following conditions were removed from the candidates, as they are likely to be noise.

Table 1: Experimental Results

	concat	infuse
MentionVec	74.71	74.80
EntityVec	75.34^{+}	75.30^{+}
ConvertedEntityVec	75.41^{+}	75.46^{++}
ConvertedBERTVec	76.06^{++}	75.86^{++}

1. There is no string inclusion relationship between the mention m and lemma of $e_i \in E(m)$.

2. The number of links from m to e_i is less than 1% of the total number of links to e_i.

There may be cases where the number of candidate entities becomes 0. In such cases, the entity identification process is not performed.

It is important to use the most accurate information possible to verify the effectiveness of entity information. Therefore, entities were manually identified with precision for some mentions. However, due to the workload, it was not feasible to manually identify entities for all mentions. Thus, manual identification was performed for the evaluation data, while automatic identification was applied to the training data. In manual identification, the task involved selecting one entity from the candidates, ranked by the number of links obtained during candidate generation. A work environment was provided where the corresponding Wikipedia pages could be referenced. In automatic identification, the entity candidate with the highest number of links obtained during candidate generation was automatically selected.

The classification accuracy was used as the evaluation metric. This metirc is calculated by

$$\frac{\text{Number of correctly classified documents}}{\text{Number of input documents}}. \quad (1)$$

5.2 Results

The experimental results are shown in Table 1 [9]. A sign test was conducted between **MentionVec** and the other methods, with "+" indicating a significant difference at the 5% significance level and "++" indicating a significant difference at the 1% significance level.

[7] https://megagonlabs.github.io/ginza/

[8] Namely, Airport,Amusement Park Archaeological Place Other Bay,Bridge,Canal,Car Stop,City,Company,Continental Region Corporation Other,Country,County,Domestic Region Earthquake,Facility Other,Facility Part,Game,Geological Region Other GOE Other,Government,GPE Other,International Organization Island,Lake,Location Other,Mountain,Museum,Occasion Other Organization Other,Park,Port,Postal Address,Pro Sports Organization Province,Public Institution,Railroad,Religious Festival,Research Institute,River,Road,School,Sea,Show Organization,Spa,Sports Facility,Sports League,Sports Organization Other Station,Theater,Tumulus,Tunnel,War,Water Route,Worship Place,Zoo.

[9] As a reference, the classification accuracy of the pure BERT document classification model without incorporating entity information was 74.33.

Table 2: Results by the number of mentions included in the document (concat)

#mentions	MentionVec	EntityVec	ConvertedEntityVec	ConvertedBERTVec	#docs (rate)
0	45.89	45.44	45.44	46.85	1,556 (22.23)
1	73.50	74.69	74.89	74.79	2,023 (28.90)
2	86.50	87.36	87.36	88.40	1,733 (24.76)
3	91.25	91.82	92.01	92.39	1,051 (15.01)
≥ 4	89.64	90.58	90.42	90.89	637 (9.10)

Table 3: Results by the number of mentions included in the document (infuse)

#mentions	MentionVec	EntityVec	ConvertedEntityVec	ConvertedBERTVec	#docs (rate)
0	44.79	45.76	45.57	45.63	1,556 (22.23)
1	73.90	74.15	75.14	75.38	2,023 (28.90)
2	86.56	87.13	86.79	87.77	1,733 (24.76)
3	91.82	92.01	91.91	92.39	1,051 (15.01)
≥ 4	90.89	91.37	91.52	91.52	637 (9.10)

From Table 1, it can be seen that, in both **concat** and **infuse** methods, the performance of the other methods improved compared to **MentionVec**, confirming that providing geographical entity information to the document geolocation model is adequate. Comparing the embedding representation acquisition methods, **ConvertedEntityVec** and **ConvertedBERTVec**, which involve conversion to prefecture names, showed higher classification accuracy than **EntityVec**. Furthermore, between the two methods involving prefecture conversion, **ConvertedBERTVec**, which acquires embedding representations through BERT, achieved better results. In this setting, it is suggested that when incorporating external knowledge into BERT, the external knowledge superficially within the input text yields better results than using embedding representations acquired independently of BERT. No apparent difference was observed between concat and infuse regarding the incorporation positions.

Next, the results for each number of mentions included in the documents are shown in Table 2 and Table 3. From these tables, it is first confirmed that the performance is significantly lower when the number of mentions is 0. This indicates that mention information is a strong clue in document geolocation. When mentions are included in the document, classification accuracy tends to improve as the number of mentions increases. However, when the number of mentions reaches four or more, the classification accuracy decreases. In documents with a relatively large number of mentions, the content often involves movement between various locations or comparisons between various locations. This complexity is likely a contributing factor to the decrease in classification accuracy.

Table 4: Results using the subset data consisting of cases that include mentions

	concat	infuse
MentionVec	82.86	83.34
EntityVec	83.82^{++}	83.69
ConvertedEntityVec	83.87^{++}	83.93^{+}
ConvertedBERTVec	84.33^{++}	84.42^{++}

Table 4 shows the classification accuracy when focusing only on the data with mentions for each method. This table summarizes the results from Table 2 and Table 3, excluding cases with zero mentions, for each method. Since most of the investigated methods showed significant improvements in classification accuracy compared to **MentionVec**, it can be said that performing entity linking and providing entity information is effective for document geolocation of documents with geographical clues.

F-score values for each prefecture class are shown in Table 5. It can be seen that the proposed methods improved performance over **MentionVec** in most prefectures. While no notable differences were observed across regional divisions, significant performance improvements were evident in prefectures with many cases, such as Tokyo, Osaka, Hokkaido, Kyoto, Kanagawa, and Fukuoka. There remain challenges in improving performance in regional areas.

Examples of classification outputs using models incorporating entity information through **ConvertedBERTVec** and **concat** are shown in Table 6. Case (**c1**) is an example where entity information led to a correct classification. In this example, the mention of "Narita" provided information about the entity "Narita International Airport," which, through the location information of "Chiba Prefecture," allowed for the correct classification. On the other hand, case (**w1**) is an example where the classification was correct with EntityVec but changed to incorrect after the conversion to prefecture names. The Zao Mountain Range is located on the border between Yamagata Prefecture and Miyagi Prefecture, but in ConvertedBERTVec, the embedding representation was acquired as Miyagi Prefecture, leading to the error. This example demonstrates cases where the conversion to prefecture names can negatively impact.

6 Conclusion

We discussed incorporating geographical entity information into the document geolocation models. The experimental results demonstrated the effectiveness of geographical entities. In particular, embedding representations that focus on entity location information were found to function effectively. Future challenges include expanding entity information from sources like Wikipedia and exploring the learning of embedding representations for entity information using frameworks such as LUKE (Yamada et al., 2020).

References

Han Bo, Cook Paul, and Baldwin Timothy. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

Lianhua Chi, Kwan Hui Lim, Nebula Alam, and Christopher J. Butler. 2016. Geolocation prediction in Twitter using location indicative words and textual features. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 227–234. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Keyur Faldu, Amit Sheth, Prashant Kikani, and Hemang Akbari. 2021. Ki-bert: Infusing knowledge context for better language and domain understanding.

Tommaso Fornaciari and Dirk Hovy. 2019. Geolocation with attention-based multitask learning models. In *Proc. 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 217–223, Hong Kong, China. Association for Computational Linguistics.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14. International Committee on Computational Linguistics (ICCL).

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*.

Sarra Hasni and Sami Faiz. 2021. Word embeddings and deep learning for location prediction: tracking coronavirus from british and american tweets. *Social Network Analysis and Mining*, 11(1):1–20.

Toi Hirakawa and Takashi Inui. 2020. Indicated deepgeo: A method for japanese document geolocation. *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2020:3Rin473–3Rin473.

Toi Hirakawa and Takashi Inui. 2022. A neural document geolocation model using geographical knowledge graph. *IPSJ Journal*, 63(12):1870–1883.

Binxuan Huang and Kathleen Carley. 2019. A hierarchical location prediction neural network for twitter user geolocation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4732–4742.

Table 5: F-score by Prefectures

					Converted		Converted		
	MentionVec		EntityVec		EntityVec		BERTVec		#docs (rate)
	concat	infuse	concat	infuse	concat	infuse	concat	infuse	
Hokkaido and Tohoku region									
Hokkaido	88.23	88.01	88.94	88.37	90.04	88.59	89.42	89.21	661 (9.44)
Aomori	73.68	81.72	82.22	76.60	77.78	76.92	79.57	83.87	50 (0.71)
Iwate	76.19	73.87	78.10	76.64	76.79	78.10	81.48	80.36	57 (0.81)
Miyagi	75.43	80.14	79.02	78.70	78.47	78.38	77.38	77.55	59 (2.27)
Akita	77.42	77.89	78.72	75.00	77.08	78.72	78.35	80.43	52 (0.74)
Yamagata	61.86	64.58	68.09	65.93	63.16	65.91	66.67	64.44	47 (0.67)
Fukushima	74.19	72.73	75.86	73.17	74.34	71.54	77.69	75.41	67 (0.96)
Kanto region									
Saitama	62.17	59.62	61.54	63.37	58.27	60.47	63.41	61.94	143 (2.04)
Chiba	75.24	73.03	72.23	69.53	72.14	72.16	70.96	71.30	274 (3.91)
Tokyo	71.73	71.85	73.20	73.41	72.73	72.41	73.45	74.10	1236 (17.66)
Kanagawa	66.15	66.35	66.35	67.07	69.01	68.90	68.04	67.61	303 (4.33)
Ibaraki	69.63	72.87	67.67	71.88	70.87	73.44	68.18	75.38	74 (1.06)
Tochigi	73.17	74.53	75.47	74.53	71.86	75.47	76.43	73.17	85 (1.21)
Gunma	69.86	70.83	70.59	67.97	66.23	70.75	72.85	76.62	87 (1.24)
Yamanashi	67.69	67.67	73.44	71.21	71.11	73.44	74.80	74.02	67 (0.96)
Nagano	81.10	80.31	80.92	77.15	80.47	78.79	82.63	79.70	129 (1.84)
Chubu region									
Niigata	70.30	74.07	72.15	74.53	70.73	73.29	74.12	75.00	88 (1.26)
Toyama	73.68	81.32	76.92	78.65	71.43	80.43	82.76	79.55	47 (0.67)
Ishikawa	81.72	82.16	82.68	82.42	82.61	82.87	81.56	82.61	97 (1.39)
Fukui	76.60	78.26	69.23	72.00	75.00	72.00	75.00	75.00	26 (0.37)
Gifu	73.28	70.07	73.13	71.64	71.64	74.24	74.07	76.12	75 (1.07)
Shizuoka	76.03	75.40	77.18	76.92	79.05	78.77	74.76	74.70	228 (3.26)
Aichi	76.37	75.19	77.07	76.49	76.90	75.08	76.90	76.03	331 (4.73)
Mie	75.41	72.88	74.80	72.73	70.87	75.21	77.69	78.63	64 (0.91)
Kinki region									
Shiga	53.70	54.72	54.72	53.70	53.06	52.83	52.83	52.83	68 (0.97)
Kyoto	72.19	71.48	71.48	73.83	74.30	71.52	74.04	72.98	321 (4.59)
Osaka	68.40	68.79	67.07	68.41	69.69	69.95	69.87	68.09	493 (7.04)
Hyogo	76.96	74.89	76.99	77.10	76.06	76.55	77.30	75.57	226 (3.23)
Nara	76.60	78.72	76.60	77.42	76.60	79.12	77.08	77.89	50 (0.71)
Wakayama	72.97	75.32	72.97	73.24	76.71	71.79	75.68	77.78	42 (0.60)
Chugoku region									
Tottori	69.23	65.38	66.67	65.38	67.92	69.23	64.29	64.29	32 (0.46)
Shimane	66.67	66.67	67.69	67.74	65.62	66.67	69.84	71.88	35 (0.50)
Okayama	70.83	69.47	64.65	65.98	63.16	66.00	74.47	70.10	52 (0.74)
Hiroshima	81.34	76.82	79.10	81.62	77.37	80.14	78.42	77.93	136 (1.94)
Yamaguchi	57.58	53.52	60.61	57.97	54.79	59.70	63.01	63.64	40 (0.57)
Shikoku region									
Tokushima	81.82	84.06	87.88	83.58	85.71	82.54	83.58	80.00	35 (0.50)
Kagawa	80.00	81.19	82.00	82.00	82.35	79.61	79.63	79.21	52 (0.74)
Ehime	80.92	83.46	82.93	83.20	81.30	82.26	81.60	81.60	65 (0.93)
Kochi	74.19	73.02	71.64	74.19	73.85	73.02	75.00	76.92	31 (0.44)
Kyushu and Okinawa region									
Fukuoka	78.89	81.14	80.27	81.63	81.51	81.73	80.00	78.50	305 (4.36)
Saga	76.60	71.11	78.26	78.26	72.34	76.60	75.56	75.56	26 (0.37)
Nagasaki	74.60	79.03	75.00	74.80	72.13	76.19	76.92	78.20	68 (0.97)
Kumamoto	68.29	75.00	70.59	69.49	75.00	74.14	76.92	76.27	69 (0.99)
Oita	75.93	77.59	80.00	77.97	82.46	81.08	78.33	83.19	60 (0.86)
Miyazaki	77.27	74.42	77.27	77.27	77.27	75.56	75.56	73.47	23 (0.33)
Kagoshima	85.37	81.99	86.42	86.59	84.15	83.23	85.19	86.96	85 (1.21)
Okinama	81.29	81.55	83.44	81.62	83.37	82.20	84.42	84.85	239 (3.41)

Table 6: Output examples

(c1)	
Input:	*Missed the flight, so now getting drunk at Narita.* （飛行機乗れなくて成田酔っ払うなう）
Output:	Chiba
Correct:	Chiba
Mention → Entity:	*Narita* → Narita International Airport (Chiba)
(w1)	
Input:	*It's snowing □□ #Zao # No wonder it's cold...* （雪だ □□ #蔵王 #寒いわけだ... ）
Output:	Miyagi
Correct:	Yamagata
Mention → Entity:	*Zao* → Zao Mountain Range (Miyagi)

Soichi Kageyama and Takashi Inui. 2022. Measuring geographic specificity for mentions and its application to document geolocation. *IPSJ Special Interest Group on Natural Language Processing (NL-253-19)*.

Jey Han Lau, Lianhua Chi, Khoi-Nguyen Tran, and Trevor Cohn. 2017. End-to-end network for twitter geolocation prediction and hashing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 744–753.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations*.

Rhea Mahajan and Vibhakar Mansotra. 2021. Predicting geolocation of tweets: Using combination of cnn and bilstm. *Data Science and Engineering*, 6(4):402–410.

Rada Mihalcea and Andras Csoma. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on Information and Knowledge Management*, pages 233–242.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2016. A simple scalable neural networks based model for geolocation prediction in twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 235–239, Osaka, Japan. The COLING 2016 Organizing Committee.

Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 7–16.

Taihei Morikuni, Mitsuo Yoshida, Masayuki Okabe, and Kyoji Umehara. 2015. Geo-location estimation of tweets with stop words detection. *IPSJ Journal (TOD)*, 8(4):16–26.

Yudai Nakamoto, Kyosuke Sezai, Koki Motokawa, Hideki Aso, and Naoaki Okazaki. 2023. Enhancing semantic understanding performance in japanese large language models using knowledge graphs. *The Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*, pages 2140–2145.

Seiji Okajima and Tomoya Iwakura. 2018a. Japanese place name disambiguation based on automatically generated training data. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

Seiji Okajima and Tomoya Iwakura. 2018b. Japanese place name disambiguation based on automatically generated training data. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing*.

Yves Scherrer and Nikola Ljubešić. 2021. Social media variety geolocation with geoBERT. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 135–140. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Proceedings of Chinese Computational Linguistics*, pages 194–206.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2016. Multiple tagging of extended named entity label to wikipedia documents. *The Proceedings of the 22nd Annual Meeting of the Association for Natural Language Processing*, pages 797–800.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454.

A Appendix

A.1 Details of the BERT Document Classification Model

The detailed settings of the BERT document classification model, which serves as the base model as described in Section 3, are explained. For the pre-trained BERT model, we used bert-base-japanese-v3 (released in May 2023) published by Tohoku University[10]. For fine-tuning the model for document geolocation, we used the training data described in Section 5. AdamW (Loshchilov and Hutter, 2017) was used as the optimization method, and Cross Entropy Loss was used as the loss function. Other hyperparameter settings are shown in Table 7. For the BERT encoder layers, only the final four layers used for classification were fine-tuned, and multiple learning rates were used based on Sun et al. (Sun et al., 2019). Since Twitter posts contain meta-information, the input to BERT was structured with the post text as the first sentence and the location information from the user's profile as the second sentence.

Table 7: Parameters of the BERT model

whole	
batch size	32
epochs	5
BERT	
maximum token size	512
lexicon size	32,768
dimensions of the hidden layer	768
dropout rate	0.1
Encoder Layer (9) learning rate	5e-6
Encoder Layer (10) learning rate	1e-5
Encoder Layer (11) learning rate	2e-5
Encoder Layer (12) learning rate	5e-5
classifier	
dimensions of the input layer	768
dimensions of the output layer	47
learning rate	5e-5

[10] https://huggingface.co/cl-tohoku/bert-base-japanese-v3