Mapping Historical Documents to Geographical Space

Takumi Hirayama
Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku
Hiroshima 731-3194 JAPAN
hirayama@ls.info.hiroshima-cu.ac.jp

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku
Hiroshima 731-3194 JAPAN
nanba@hiroshima-cu.ac.jp

Toshiyuki Takezawa
Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku
Hiroshima 731-3194 JAPAN
takezawa@hiroshima-cu.ac.jp

ABSTRACT

Geotagging is the process of recognizing place and facility names in a document, and assigning each set of latitude and longitude values. In the latter step, an external geographic database, which contains pairs of place/facility names and latitude/longitude values, is used. However, if former place/facility names are used in a historical document, it is impossible to assign latitude and longitude values to them, even though their current names are listed in the database. Furthermore, if there are multiple identical place/facility names in the geographical database, we will have to choose the correct one. In this paper, we propose a method to construct a database that contains current and former place/facility name pairs. We applied a machine learning-based information extraction method to some text corpora, and automatically extracted current and former place/facility name pairs. We also propose a method that disambiguates the same place/facility names. We conducted some experiments to confirm the effectiveness of our method.

CCS Concepts

- Information systems~Geographic information systems
- Computing methodologies~Natural language processing
- Computing methodologies~Information extraction

Keywords

Mapping; Information Extraction; Geotagging; Place Name Disambiguation

1. INTRODUCTION

In this paper, we construct a system that maps historical documents to geographical space. Our task is considered as a kind of geotagging, which maps documents to geographical space. In general, geotagging consists of two steps: (1) recognizing place and facility names in a document, and (2) assigning each latitude and longitude values. Step 2 has the following two problems. The first problem is the ambiguity of place/facility names. If the same place/facility names extracted in Step 1 are used in different areas, we will have to disambiguate the names. The second problem is, if former place/facility names are used in a historical document, it is impossible to assign latitude and longitude values to them, even though their current names are listed in the database.

For the first problem, we propose a method that disambiguates facility/place names using the contextual information of each name in a document. For the second problem, we construct a database that contains current and former place/facility name pairs. We applied a machine learning-based information extraction method to some text corpora, and automatically extracted current and former place/facility name pairs.

The remainder of the paper is structured as follows. In Section 2, we discuss related work. In Section 3, we propose a geotagging method and a place/facility name disambiguation method. We conducted some experiments to confirm the effectiveness of our method. The experimental results are reported in Section4. Finally, Section 5 concludes our work.

2. RELATED WORK

Lieberman et al. [3] proposed a method that identifies place names in a document using manually created rules. We use a machine learning-based named entity recognizer CaboCha.

Zhao et al. [6] proposed an algorithm, GeoRank, for the disambiguation of place/facility name. The basic idea of this algorithm is based on PageRank [1], which is well-known as a ranking algorithm for Web search. A candidate for an ambiguous place/facility name can give more evidence to the one near to it in a document, and a location can give more evidence to the one near to it in the geographic context. Therefore, they constructed a matrix involving all locations, whose values are scores of each place/facility of each candidate voted by other ones that belong different candidates. Our method for the disambiguation also employs the voting, but is simpler and faster than GeoRank, because we did not use PageRank.

3. GEOTAGGING

3.1 Construction of a Former/Current Place Name Dictionary

3.1.1 Basic Strategy

In some documents, both former and current place/facility name pairs are described in the same sentence. The following sentence is such an example.

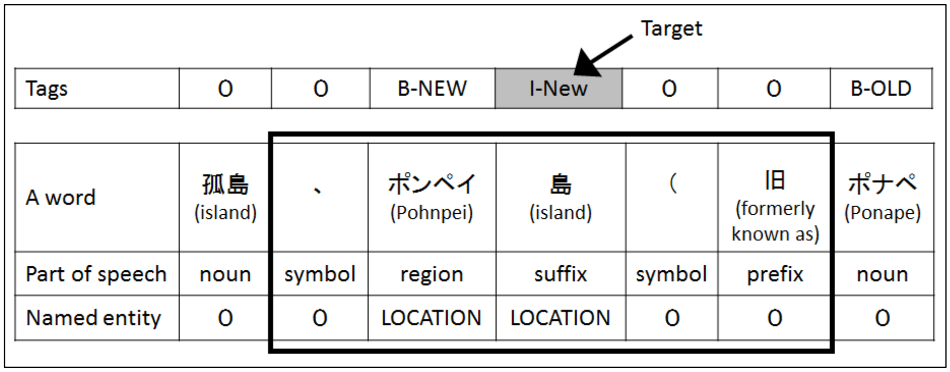
[Original]

毎年恒例ミクロネシアツアー。今年は太平洋の孤島、ポンペイ島（*旧*ポナペ）です。

[Translation]

This year, Micronesia tour visits Pohnpei Island *(formerly known as* Ponape Island) in the Pacific Ocean.

MOBIQUITOUS '16 Adjunct Proceedings, November 28-December 01, 2016, Hiroshima, Japan
© 2016 ACM. ISBN 978-1-4503-4759-4/16/11…$15.00
DOI: http://dx.doi.org/10.1145/3004010.3004028



Tags	O	O	B-NEW	I-New	O	O	B-OLD
A word	孤島 (island)	、	ポンペイ (Pohnpei)	島 (island)	(	旧 (formerly known as)	ポナペ (Ponape)
Part of speech	noun	symbol	region	suffix	symbol	prefix	noun
Named entity	O	O	LOCATION	LOCATION	O	O	O

In this sentence, the underlined and dashed line phrases indicate current and former place names, respectively. Therefore, we can extract the current and former place name pairs from this sentence using the following linguistic pattern.

[Original]

“[current name]（旧[former name]”

[Translation]

“[current name] (formerly known as [former name]”

However, there are several cases that do not follow the above pattern, such as shown in the following examples.

(Example 1)

[Original]

埼玉５区：さいたま市(旧浦和市、大宮市)

[Translation]

Saitama district 5: Saitama City (Formerly known as Urawa City and Omiya City)

(Example 2)

[Original]

特定商取引に関する法律（旧通販法）

[Translation]

Act on Specified Commercial Transactions (formerly known as Mail order law)

In example 1, one current and two former names appear in the same sentence. In this case, if we apply the above linguistic pattern to this sentence, we cannot extract the latter former name (“大宮市” (Omiya City)). In example 2, the names given before and after the expression of the linguistic pattern “(旧” (formerly known as) are not place names. Therefore, we apply a machine learning-based information extraction approach, instead of simply applying the linguistic pattern. We define this task as a sequence-labeling problem, and apply conditional random field (CRF) as a machine-learning method, which is widely used in various natural language processing tasks, such as part of speech tagging [2] and named entity recognition [4]. We will describe our approach in detail in the following section.

3.1.2 Automatic Annotation of NEW/OLD Tags

We prepare sentences that contain an expression “（旧” ((formerly known as), and manually annotate these sentences with NEW and OLD tags. Here, NEW and OLD tags indicate current and former location/facility names, respectively. The following are examples of manually tagged sentences.

[Original]

毎年恒例ミクロネシアツアー。今年は太平洋の孤島、<NEW>ポンペイ島</NEW>（旧<OLD>ポナペ</OLD>）です。

[Translation]

This year, the Micronesia tour visits <NEW>Pohnpei Island</NEW> (formerly known as <OLD>Ponape Island</OLD>) in the Pacific Ocean.

We used these tagged sentences for training and evaluation purposes using CRF. To employ CRF, we converted texts into morphological features as 1-4 grams of word, parts of speech, and named entities. To obtain these features, we used a Japanese dependency parser CaboCha.[1] CaboCha can annotate several named entity tags: “LOCATION,” “ORGANIZATION,” “PERSON,” “DATE,” “TIME,” “ARTIFACT,” “PERCENT,” and “MONEY.” Among them, we used “LOCATION,” “ARTIFACT,” and “ORGANIZATION” for our task. Figure 1 shows how NEW and OLD tags are used in annotating using CRF. In this figure, we do not show 2-4 gram features because of the limitation of the pages. In Figure 1, we attempt to identify a tag for a target word “島” (island). To identify this tag, we use features appearing before and after this target word.

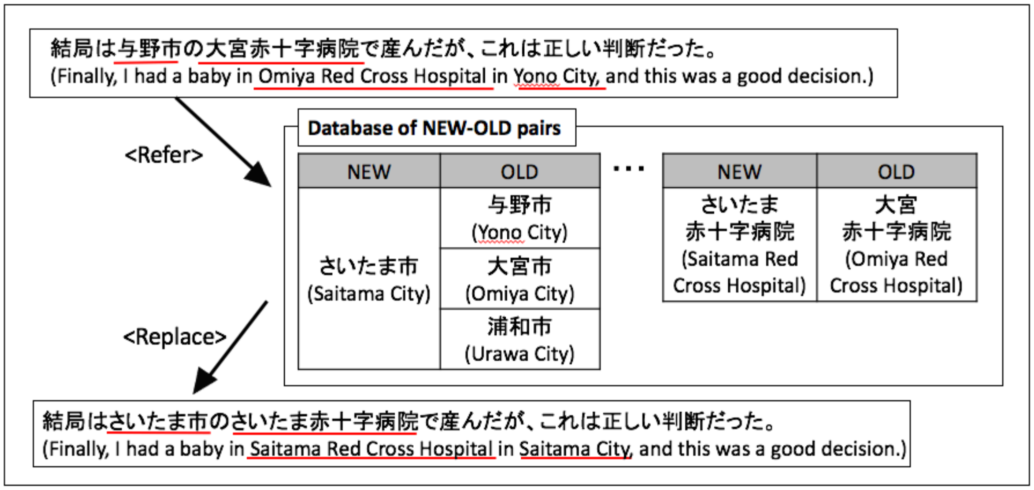
3.2 Application of a Former/Current Place Name Dictionary to Documents

Our system identifies all former place/facility names in a historical document using CaboCha and replaces them with the corresponding current names. We explain this procedure using an example in Figure 2. First, our system identifies all place/facility names in a document. As a result, “与野市” (Yono City) and “大宮赤十字病院” (Omiya Red Cross Hospital) were identified. Second, we compare these names with all OLD item records in the dictionary in Section 3.1. If we find the same record, then we replace the name in the document with the corresponding current name. In case of the example in Figure 2, “与野市” (Yono City) and “大宮赤十字病院” (Omiya Red Cross Hospital) are replaced with “さいたま市” (Saitama City) and “さいたま赤十字病院” (Saitama Red Cross Hospital), respectively.

3.3 Disambiguation of Place/Facility Names

Geotagging is performed in this step. For each place/facility name in the document, we assign latitude/longitude values using several gazetteers. As gazetteers for geotagging, we used 1,957 prefecture and municipality names provided by the Geospatial Information

[1] https://taku910.github.io/cabocha/



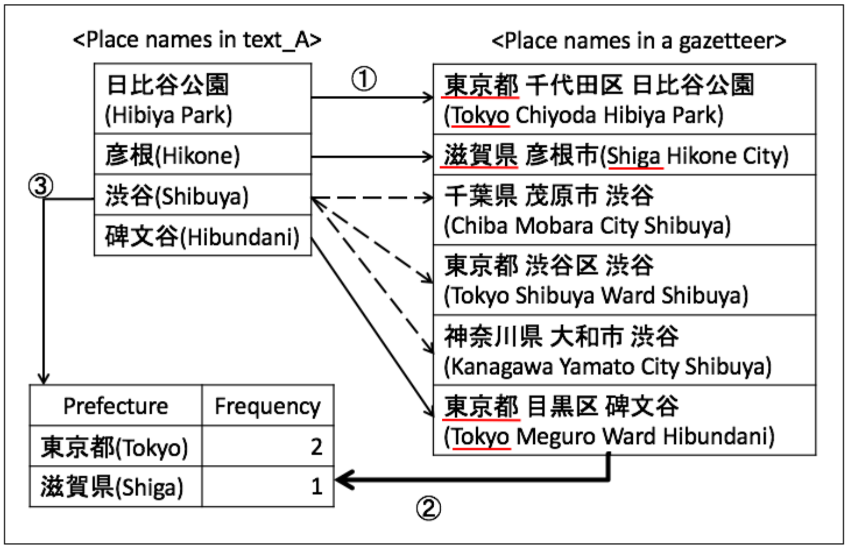
結局は与野市の大宮赤十字病院で産んだが、これは正しい判断だった。
(Finally, I had a baby in Omiya Red Cross Hospital in Yono City, and this was a good decision.)

NEW	OLD
さいたま市 (Saitama City)	与野市 (Yono City)
	大宮市 (Omiya City)
	浦和市 (Urawa City)

NEW	OLD
さいたま 赤十字病院 (Saitama Red Cross Hospital)	大宮 赤十字病院 (Omiya Red Cross Hospital)

結局はさいたま市のさいたま赤十字病院で産んだが、これは正しい判断だった。
(Finally, I had a baby in Saitama Red Cross Hospital in Saitama City, and this was a good decision.)

Figure 2. Replacing former place/facility names with current place/facility names



<Place names in text_A>
<Place names in a gazetteer>
日比谷公園 (Hibiya Park)
彦根(Hikone)
渋谷(Shibuya)
碑文谷(Hibundani)
東京都 千代田区 日比谷公園 (Tokyo Chiyoda Hibiya Park)
滋賀県 彦根市(Shiga Hikone City)
千葉県 茂原市 渋谷 (Chiba Mobara City Shibuya)
東京都 渋谷区 渋谷 (Tokyo Shibuya Ward Shibuya)
神奈川県 大和市 渋谷 (Kanagawa Yamato City Shibuya)
東京都 目黒区 碑文谷 (Tokyo Meguro Ward Hibundani)
①
②
③

Prefecture	Frequency
東京都(Tokyo)	2
滋賀県(Shiga)	1

Figure 3. Disambiguation of place/facility names

Authority of Japan,[2] 117,061 town/area names provided by Gengo-Shigen-Kyokai,[3] and 44,930 entity names, such as stations and mountains, obtained sfrom Wikipedia.

If there are multiple candidates for a place/facility name, that is, multiple records include the same names as in gazetteers, we will have to choose the correct record from these candidates. In the following, we explain our method of disambiguating place/facility names using Figure 3. Our method consists of the three steps. In the first step, our method assigns latitude/longitude values to nonambiguous place/facility names in a document. In Figure 3, "日比谷公園" (Hibiya Park) , "彦根" (Hikone), and "碑文谷" (Hibundani) are nonambiguous names. In the second step, we count the number of prefectures for each record according to nonambiguous place/facility name. The purpose of this step is to estimate the geographical area of the document. As "渋谷" (Shibuya) appears in Chiba, Tokyo, and Kanagawa prefectures, we consider that "渋谷" (Shibuya) is an ambiguous phrase, and eliminate it from the prefecture count. As a result, we can estimate that geographical area of this document is "東京" (Tokyo), because its frequency 2 is higher than that of "滋賀" (Shiga). In the final step, we chose one candidate with a geographical area that matches the prefecture in the previous step. In Figure 5, we chose "東京都渋谷区渋谷" (Shibuya, Shibuya Ward, Tokyo) as the corresponding record.

4. EXPERIMENTS

4.1 Extraction of Former/Current Name Pairs

4.1.1 Data Sets and Experimental Settings

We randomly selected 6,914 sentences containing the expression "（旧" ("(formerly known as") in the NTCIR-5 Web database [5] dataset and manually labelled these sentences with 3,958 NEW tags and 4,002 OLD tags. We employed CRF as a machine-learning technique and performed a fivefold cross-validation using the recall, precision, and F-measure evaluation measures.

[2] http://www.gsi.go.jp/ENGLISH/index.html

[3] http://www.gsk.or.jp/en/

We compared these results with the results from a baseline method that extracts phrases appearing before and after a linguistic pattern "（旧" as NEW/OLD pairs.

4.1.2 *Experimental Results and Discussion*

Table 1 shows the results of the experiment. The table shows that all values in our method exceed 0.8. In particular, the precision scores are approximately 0.9. In addition, our method obtained a high F-measure compared with the baseline method. Our method could not extract some NEW/OLD pairs, such as general store name pair "プラザ" (PLAZA) and "ソニプラ" (Sonipla). In case of this pair, our method could not assign either NEW or OLD tags. Another typical error in our method is "二葉館" (Futaba Museum) and "川上貞奴邸" (Kawakami Sadayakko's Residence). In this case, "文化のみち二葉館" (Cultural Path Futaba Museum) should be extracted instead of "二葉館" (Futaba Museum).

Table 1. Evaluation results for the NEW-OLD pair extraction.

Method	Tag	Precision	Recall	F-measure
Our method	NEW	0.889	0.805	0.845
	OLD	0.880	0.823	0.851
Baseline	NEW	0.572	1.000	0.728
	OLD	0.572	0.989	0.725

We applied our method to the following three corpora (Table 2) and constructed a former/current name pairs dictionary. NEW and OLD tags are annotated to the sentences in each of these corpora. Among the tagged phrases, we eliminated pairs, whose frequencies are low. We also eliminated very long or very short phrases. The statistics of the results are shown in Table 3.

Table 2. Statistics for three corpora

Corpus	Number of documents	Size	Publication years
NTCIR5Web	10 million	1.3 TB (uncompressed)	Crawled in 2004
ClueWeb09	6.7 million	500 GB (compressed)	Crawled in 2009
Yomiuri newspaper	5.4 million	9.1 GB	1993-2012

Table 3. Statistics of three corpora

Corpus	Number of NEW tags	Number of OLD tags
NTCIR5Web	363,888	348,055
ClueWeb09	745,913	831,412
Yomiuri newspaper	42,505	50,138

Tables 4, 5, and 6 show some extraction results, former/current name pairs with their frequencies, from NTCIR5Web, ClueWeb09, and the Yomiuri newspaper databases, respectively. Shaded phrases in Table 5 indicate mistakenly extracted current/former name pairs.

Table 4. Extraction results from NTCIR5Web

Freq	Current (NEW)	Former (OLD)
16,061	ジェネオン (Geneon)	パイオニア LDC (Pioneer LDC)
15,834	HP	コンパック (Compaq)
1,429	グレープシティ株式会社 (GrapeCity Inc.)	文化オリエント株式会社 (Bunka Orient Corp.)
1,094	潮来市 (Itako City)	牛堀町 (Ushibori Town)
1,064	幌内太駅 (Horonaibuto Station)	三笠駅 (Mikasa Station)
1,056	西日本工業倶楽部 (The Industry Club of West Japan)	松本邸 (Matsumoto's Residence)
1,023	都筑民家園 (Tsuzuki Minka-en)	長沢家住宅 (Nagasawa's Residence)

Table 5. Extraction results from ClueWeb09

Freq	Current (NEW)	Former (OLD)
1,438	コンゴ (Congo)	ザイール (Zaire)
1,107	HP	コンパック (Compaq)
1,068	北栄町 (Hokuei Town)	大栄町 (Daiei Town)
836	住宅金融支援機構 (Japan Housing Finance Agency)	住宅金融公庫 (Government Housing Loan Corporation of Japan)
817	ロシア(Russia)	ソビエト連邦 (the Union of Soviet Socialist Republics)
751	セ映画館(Se-movie theater)	相鉄ムービル等 (Sotetsu-mubiru etc.)
588	静岡県伊東市(Ito City, Shizuoka Prefecture)	国伊豆国(Province Izu Province)

Table 6. Extraction results from Yomiuri newspaper

Freq	Current (NEW)	Former (OLD)
264	中国東北部(Northeast China)	満州(Manchuria)
218	コンゴ民主共和国 (Democratic Republic of the Congo)	ザイール(Zaire)
121	新生銀行(Shinsei Bank, Limited)	日本長期信用銀行(The Long-Term Credit Bank of Japan)
68	ロシア(Russia)	ソ連(USSR)
61	都市再生機構(Urban Renaissance Agency)	都市基盤整備公団 (Urban Development Corporation)
58	ミャンマー(Myanmar)	ビルマ(Burma)
56	イオン(AEON)	ジャスコ(JUSCO)

As an overall trend, we obtained more computer company name pairs, such as “HP” and “コンパック” (Compaq), from NTCIR5Web and ClueWeb09 databases than from the Yomiuri newspaper database, because more computer-related information is available on the Web than from news articles. The Yomiuri newspaper database extracted more administrative corporations than NTCIR5Web and ClueWeb09.

We compared the pairs from NTCIR5Web with those from ClueWeb09, and found that there are apparent differences between them. For example, we can find a pair of “住宅金融支援機構” (Japan Housing Finance Agency) and “住宅金融公庫” (Government Housing Loan Corporation of Japan) in Table 5. This pair did not appear in NTCIR5Web, because “住宅金融支援機構” (Japan Housing Finance Agency) was established in 2007 and this organization did not exist when NTCIR5Web database was created. This result shows that we can expect to extract former/current name pairs for each year if we prepare only document corpora that were published in that year. Using this data, we might obtain the history of a particular place/facility name.

4.2 Geotagging

In this section, we describe the matching technique experiments in geotagging place names. Section 4.2.1 describes the experimental method. Section 4.2.2 provides the experimental results and discussion.

4.2.1 Data Sets and Experimental Settings

Fifty randomly selected articles published in 1993, 1998, 2003, 2008, and 2012 in Yomiuri newspaper were used. We obtained 177 place/facility names using CaboCha. Twenty ambiguous names were identified. We manually disambiguated these ambiguous names and used them for our experiment. We used the precision evaluation measure. A baseline method was prepared, which randomly selected candidates to confirm the validity of our proposed method.

4.2.2 Experimental Results and Discussion

Table 7 shows the experimental results, which demonstrate that our method outperformed the baseline method.

Table 7. Evaluation results for place/facility name disambiguation.

Year	Number of names	Number of ambiguous names	Baseline	Our method
1993	22	6	0.727	0.909
1998	41	2	0.951	0.951
2003	31	1	0.968	1.000
2008	31	3	0.900	0.935
2012	52	8	0.840	0.942
Average	35.4	4	0.877	0.947

Unfortunately, only four of the 177 were ambiguous names and we could not confirm the effectiveness of our former/current place/facility name dictionary in this experiment. We will apply our method to other historical documents in future, such as the ancient documents held by the National Archives of Japan.[4]

5. CONCLUSIONS

In this paper, we proposed a method to map historical documents to geographical spaces. A dictionary of former/current place/facility names was constructed from several corpora using a CRF-based method. We also proposed a method for disambiguating place/facility names. We conducted some experiments, and confirmed that our method obtained F-values of approximately 0.85 for former/current names. We also confirmed that our method obtained a precision score of 0.949 for disambiguation of place/facility names.

In future, we will apply our method to English documents, and extract former/current name pairs from them. As we did not use any features restricted to Japanese, we believe that it is easy to apply our method to documents written in any languages.

6. ACKNOWLEDGMENTS

This study was carried out under support of Ministry of Internal Affairs and Communications' SCOPE (Strategic Information and Communications R&D Promotion Programme).

7. REFERENCES

[1] Brin, S. and Page, R., The Anatomy of a Large-scale Hyper Textual Web Search Engine, In Proceedings of World Wide Web, pp.107-117, 1998.

[2] Kudo, T., Yamamoto, K., and Matsumoto, Y., Applying Conditional Random Fields to Japanese Morphological Analysis, In Proceedings of the 2004 Conference on Empirical Methods on Natural Language Processing, 2004.

[3] Lieberman, M.D., Samet, H., and Sankaranayananan, J., Geotagging: Using Proximity, Sibling, and Prominence Clues to Understand Comma Groups, In Proceedings of GIR’ 10, 2010.

[4] McCallum, A. and Li, W., Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons, In Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003, pp. 188-191, 2003.

[5] Oyama, K., Takaku, M., Ishikawa, H., Aizawa, A., and Yamana, H., Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2 (Navi-2), In Proceedings of NTCIR-5 Workshop, 2005.

[6] Zhao, J., Jin, P., Zhan, Q., and Wen, R., Exploiting Location Information for Web Search, Computers in Human Behavior, 2013. http://dx.doi.org/10.1016/j.chb.2013.04.023

[4] http://www.archives.go.jp/