Extraction and disambiguation of name of place from tourism blogs

Nakatoh, Tetsuya
Research Institute for Information Technology, Kyushu University

Yin, Chengjiu
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings - 1st ACIS International Symposium on Software and Network Engineering, SSNE 2011, pp.73-78, 2011-12-01
バージョン：
権利関係：

Extraction and Disambiguation of Name of Place from Tourism Blogs

Tetsuya Nakatoh, Chengjiu Yin, Sachio Hirokawa
Research Institute for Information Technology, Kyushu University
{nakatoh,yin,hiroakwa}@cc.kyushu-u.ac.jp

***Abstract*—By development of the Internet in recent years, tourism portal sites and blog articles about tourism increased on WWW. Acquisition of various tourism information became easy. When gathering and classifying the information automatically from blog articles, it is not easy to decide automatically place names used as the key. In this paper, we propose a method of extracting place names from blog articles automatically. Moreover, we also tried disambiguation of a place name.**

I. Introduction

The information on the Internet is continuing increasing every day. Thanks to the development of search technology, anyone can obtain various information in large quantities nowadays. Tourism information is not an exception. Most people use the Internet to obtain the information concerning the destination and stay before they actually going out for sightseeing. These tourism information are available in

(a) Tourism portal sites,
(b) Web pages of service providers and
(c) users' Blog pages.

For example, official information with respect to the accommodation and hotels are available at travel agent's reservation service site. Many people write their comments and opinion in so-called "word-of-mouth" site where we can use others' evaluation for choosing a hotel. However, writing a critical comment is not easy. In fact, the ratio of negative opinions are very small compared to the positive ones. Anyway, perfect neutrality is not guaranteed.

There are a large number of blog articles where general users write individual experiences, travel records and opinions. These articles are not official ones but can be reliable in a sense that they have no influence from organizations. The blog articles cover a wide range of topics. It is often the case that a user writes where they visited and what sort of food they enjoyed, together with their comment on the hotel where they stayed. These information cannot be found at a hotel site.

However, as compared with an official site, neither a detailed regional name nor an institution name are clear in many cases. To use these personal opinion, we first need to extract the exact place names and we need to confirm the real address of the spots. It is relatively easy for human to understand the place name, since we have a lot of background knowledge and we can grasp the context. However, automation of the processing is not easy. The authors have been working concerning to search engine of tourism information. The present paper reports a trial of place name extraction from tourism blogs and disambiguation of location names.

II. Related work

There are many researches in collecting and extracting valuable information from Web pages or news paper articles. For example, [12] tried to define tourism information by extracting keywords from WWW documents. They proposed a key map which visualizes co-occurrence relation, and showed the concrete key map obtained from the search results with respect to the keywords of tourism in Hokkaido or Okinawa. [11] proposed a natural language interface for the sightseeing tour search engine. In order to offer the travel plan according to an individual's liking, [2] considered that collecting and updating sightseeing information are crucial. They realized information extraction agent (IE) and information clustering agent (IC) as add-on function to the tourism recommendation system. They demonstrated the effectiveness by providing the names of services and location, the price, the time, and the period which they extracted from Web pages using patterns.

The targets of these researches are Web pages offered by the travel agents or the related organization. The objects of the present paper on the other hand are various blog pages written by general users and there is no common pattern in those pages. Many people write their experience and opinion on their blogs.

Attentions are being paid for the analysis and practical use of these blogs. [4] proposed a method to extract the keyword from tourism blogs that characterize the area. [9] succeeded to choose feature sentences using sentiment information and clustered the sentences depending on semantic attributes of part of speech and the characteristic nouns with tourism. [13] analyzed the difference between the information provided by the travel agent and those by blogs concerning tourism. [14], [15] proposed the methods to extract the events of the regions.

The name of places are typical example of the named entities for which there have been many researches. [7] analyzed characteristic occurrences of tag patterns to extract the names of place and the names of events from tourism related Web pages provided by the organizations. [10] focused on the periodicity of touristic events found in news papers

to discover such events. Both used the structural pattern of the documents which cannot be expected with respect to blogs written by general users. [6] applied a machine learning method to extract the pairs of the place name and the souvenir from tagged corpus. However, the specification of the regional names are out of scope. Indeed, they uses the parser Cabocha[1] to designate the regional names.

In order to utilize the blogs as sightseeing information, extraction and word sense disambiguation of the name of a place are an important issue which cannot be bypassed. As we described in this section, there is various research, but there are no comprehensive and a uniform valuation method of polysemy discernment.

In this paper, the name of a place which appears in four large-scale place name databases was made into the candidate, and the tourist resort name list with high comprehensibility was made from limiting only to what has appeared in the context of sightseeing.

In order to discovery and evaluate ambiguity of place name, it is necessary to classify the contexts in which the same name is used as different places. We paid attention to the fact that the names of places coincide with specific verbs and particles, when they are used in relation to sightseeing. We used the concept graph [1], [5] as visualization of polysemy of the name of a places.

III. Dependency structure analysis

There are two methods considered to identify the name of a place in a text. The first method identifies the character string to describe the name of a place using a name of a place dictionary. Another method determines the name of a place portion by making other surrounding words a key. The first method will solve the problem, if we had a perfect dictionary of place names. However, it is not easy task to construct such a dictionary. The names of places do not stay. New names of places will appear according to the merger of cities, towns and villages. There are many abbreviated names used only in local. The name of an institution is often used as the name of place. There are many newly opened institute. And some institutes change their names. The official name of the baseball stadium in Fukuoka is Yahoo dome. But, many local people refer the stadium as Fukuoka dome as it was used to before the owner changed.

Imagine that we read a sentence which contain a place name. If the place is familiar to us, we recognize the name as a place name. Even if we do not know the place, we can guess that a name represents a place from the context where the word appears. For example, we can recognize the place names “Fukuoka” and “Hakomatsu” from the sentences “I came from Fukuoka” and “I have been to Hakomatsu, yesterday”. Such words that appear in particular contexts and in dependence structures can be recognized as place names even if we do not know the words.

The combination of the two identification methods of place names will improve the precision and the recall of identified names. Both methods have the similar difficulties that we should know some knowledge before we apply them. The first method requires a list place names. The second method requires a list of dependence pattern in which place names appear. The two knowledge are related each other in such a way that we can learn the contextual patters where the known place names appear and that we can extract the place names from the instances of contextual patterns. This approach can be considered as an application of bootstrapping to the place names and the contexts of place names.

We propose a method to use the dependency structure analysis for identification of place names. We focus on the case-marking particles that specify the target location of our behavior in Japanese. A case-making particle designates the place where an action occurs or specifies the start point and the end point. We can identify the words as the place names if they appear as targets of a case-making particle. We can increase our list of place names by this inference. A case-marking particle occurs as a patter of “Noun + Case-marking particle +Verb”. However, there is no such a list of patterns known to identify place names.

The present paper is an initial step to construct such a list of patterns and a list place names. The key data of our bootstrapping method are the list of triples of place name, case-marking particle and verb. We collected 7,917,385 blog articles and obtained 132,343,454 sentences to which we analyzed dependence structure using CaboCha[2]. As the result, we obtained 19,842,569 triples of noun, case-marking particle and verb. Table I shows the number of occurrences of the triples of “Tenjin” with top 10 verbs and top 6 case-marking particles. The number after each word represents the number of occurrences of the word.

The most frequent words used to describe behavior with place names is “iku (go)”. “aru (be)” is the second verb whose occurrences are limited with the case-makring particle “ni (for)”. Other verbs have the similar preference of the particular case-makring particle.

The top 5 case-makring particles cover 90% of all occurences. The particle “de (to)” has a distinct character such that it is used in the context which does not refer to movement. Other particles are usually used with motion or movement. The particle “de (at)” is used to specify tools, materials and cause other than location.

IV. Extraction of place Name from blog

The blog articles on a specific area or a place can be obtained simply by sending the name of the place as a

[1] http://code.google.com/p/cabocha/

[2]

Table I
TOP 10 VERB*PARTICLE PAIRS FOR "TENJIN"

	ni (for)	made (till)	e (toward)	de (at)	kara (from)	wo	no (of)	mo (too)	wa	to (and)
	432	97	86	82	66	21	9	7	7	5
iku (go) 192	124	37	18	5	0	1	0	0	1	0
aru (be) 83	73	0	1	3	0	0	1	1	2	1
deru (out) 43	26	9	7	0	0	0	0	0	0	1
aruku (walk) 34	4	12	2	1	9	6	0	0	0	0
dekaeru (depart) 31	16	7	7	0	0	0	0	0	1	0
modoru (return) 28	12	5	7	0	4	0	0	0	0	0
okonau (do) 25	3	0	12	6	2	0	0	0	0	1
suru (do) 24	0	1	1	14	0	4	0	0	1	0
kuru (come) 23	16	6	1	0	0	0	0	0	0	0
tsuku (arrive) 22	20	0	1	0	0	0	0	0	0	0

query word for a search engine. However, the article is not necessarily the correct one as the user is looking for. It is necessary to read the text of search results and to judge individually. There are many differences of place names in the dictionaries and in the blog articles. In this paper, we use the following three methods to compile an exhaustive list of tourism related place names.

(a) Limitation of the sightseeing context by a dependency analysis
(b) Integration of place name dictionaries
(c) Redundant registration of the name of a place

A. Limitation of the sightseeing context by a dependency analysis

It is often the case that the name of a place which appears in the text of a blog article is either a starting point of an act, a terminal point, or a destination. We apply a dependency analysis to the text of a blog article to obtain the dependency pairs of the noun and verb accompanied by a particle. The list of candidate names of place is generated using the pairs.

The location related case-marking particles are shown in TableII with their functions. We exclude the particles "wo" and "de", since they are often used not only in place but in other situations. We consider only the verbs, in the dependence analysis, to "go" and "come" to restrict the target to be place. We construct the list of the place names as the dependent nouns to which "go" and "come" depend.

Table II
PARTICLES FOR PLACE AND ITS FUNCTION

Particle	Function
wo	place of movement, start point
ni (for)	destination, reaching point, location
e (toward)	direction
de (at)	place of operation
kara (from)	starting point
made (till)	end of movement

B. Integration of place name dictionaries

Some general nouns, such as "asobi" (play) or "issho" (together) might appear with "go" and "come". To remove these non-place nouns, we use the words in the place name dictionaries.

C. Redundant registration of the name of a place

The name of a place as administration units, such as "Fukuoka city", is contained in expression of the existing name of a place dictionary in many cases. However, the portion of "Fukuoka" is used as the name of a place in the usual conversation and in the blog articles. To guarantee the flexibility, we adapt both forms of "Fukuoka city" and "Fukuoka" in the index of place names.

V. BLOGS IN KYUSHU AREA

We collected 7,917,385 blog articles related to Kyushu area using by blog search. Articles other than tourism information are also contained in the blogs collected this time. Only the text in the "body" tag was extracted from HTML of the collected blogs. The extracted text consisted of 5,081,636,114 bytes, and contains 132,343,454 lines.

A. Place name dictionaries

The following four databases were used to compile a place name dictionary for analysis.

Postal-services incorporated company exhibits postal code number data [3]. 122,999 addresses with postal codes in Japan are included in CSV-form. The place name dictionary of ATOK contains 101,454 place names by individual possessions [4]. "Japan place name list 2007" is owned by Japanese government [5]. It contains the longitude and latitude, reading, and roman alphabet notation of each place. It contains the names of mountain, river, cape and peak as well. However, the total number of items is small. There are many place

[3]http://www.post.japanpost.jp/zipcode/download.html
[4]http://homepage3.nifty.com/t-weekly/
[5]http://www.gsi.go.jp/kihonjohochousa/gazetteer.html

Table III
LIST OF PLACE NAME DATABASES

database	the number of places	source
zip-code	122,999	Japan Post
ATOK	101,454	individual possessions
Japan place name list 2007	3,867	Geospatial Information Authority of Japan
NAIST-jdic	67,325	NAIST

names which cannot be listed in the database. NAIST-jdic[6] is the succeeding dictionary of IPAdic which are used as the dictionary for Japanese morphological-analysis tools ChaSen and MeCab by Nara Institute of Science and Technology (NAIST). The dictionary contains 280,562 words and 67,325 words as place name.

There are two purposes of using place name dictionaries. The first one is to extract short-names of official full names. We do not use the full name of place, for example "Akasaka, Chuo-ku, Fukuoka city, Fukuoka prefecture". We simply use the short name of "Akasaka". The place name in blogs is used to refer to the place, and not used to specify the exact geospatial information of the place. Even if there might be ambiguity and we would care, we use short name of "Akasaka, Fukuoka" and "Akasaka, Tokyo" to make distinction. We mainly used the ATOK dictionary and used other dictionaries auxiliary.

The second purpose is to the designate geographic information for the place name. The short name such as "Akasaka" is not enough for this purpose. Instead, the full name, such as "Akasaka, Minato-ku, Tokyo", which determines the location, is necessary. The zip-code dictionary and Japan place name list 2007 are appropriate for this purpose, among which we choose the zip-code in the present paper as the main dictionary.

B. Obtained place names

It is difficult to compare all the words in the 7,917,385 blog articles with more than 100,000 place names in the dictionaries. The place name in a blog can be a segment of some place name in the dictionary, which causes much more difficulty in matching. Another difficulty is that some place names are used as the names of person.

To overcome these difficulties, we use the dependency analysis of words in a sentence. We used the tool Cabocha to extract 527,471 occurrences of dependent pairs from the blog argicles. We use the pattern of "Noun"+"Case-marking particle"+"Verb", where we choose "NI (for)","E (toward)","KARA (from)" and "MADE (till)" as the case-marking particle, "IKU (go)" and "KURU (come)" as the verbs. From these occurrences, we obtained 90,057 N-P-V pairs which contained 527,471 nouns. However, we cannot simply accept these nouns as place names. In fact, the most frequent pairs, such as "ASOBI NI KURU (come to play)" and "ISSHO NI IKU (go together)", tend to contain non-place names. So, we used the list of place names that we constructed from the dictionaries as in the previous subsection, to remove non-place names and obtained 943 place names in Kyushu area. This number might look small. However, this number means that we obtained 120 place names for each 7 prefecture in Kyushu area. So, we think that the number place names is reasonable to be analyzed even if they are not large.

[6]http://sourceforge.jp/projects/naist-jdic/

VI. DISAMBIGUATION OF PLACE NAME

A place name may be used to refer to the different locations in the different blog articles. In fact, there are 181 locations which are called as "TENJIN" around Japan. Evaluation of relative location using MBR (The Minimum Bounding Rectangles) is known as a method to distinct the locations of the same place name [8]. However, the co-occurrence of the two place name does not always imply one location with the two name. For example, a blog containing "KAGOSHIMA" (the most south prefecture in Kyushu Area) and "TENJIN" might be a story about the "TENJIN" town of "KANAYA" city in "KAGOSHIMA" prefecture, or a story about a express buss from "KAGOSHIMA" to the "TENJIN" town of "FUKUOKA" city in "FUKUOKA" prefecture which locates north in Kyushu area.

In this paper, we use the notion of "classification network (CN)" [5], to formulate a context of a place name in a set of documents. It is a directed graph of words that appear in a set of document to be analyzed. A frequent word is drawn in the left of an edge and less frequent word is drawn in the right of the edge. The notion can be considered as an extension of the formal concept [1]. We generated the classification network of words given a set of blog articles, where only the place names are considered.

Figure 1 is the classification network of the 25,113 blog articles that contain "TENJIN". Most frequent 41 place names are shown. If we focus on the roots of the graph, we see 10 sub-graphs as shown in Table IV.

The subgraphs of 1,2,4 and 6.3 are easy to understand the place names as the locations, since the place names in the lower position represent wider region. In the graphs of 3,6 and 6.1, we can see many names of region related to "TENJIN". However, the relative position of words in the graph is not clear compare to the first group (1,2,4 and 6.3),

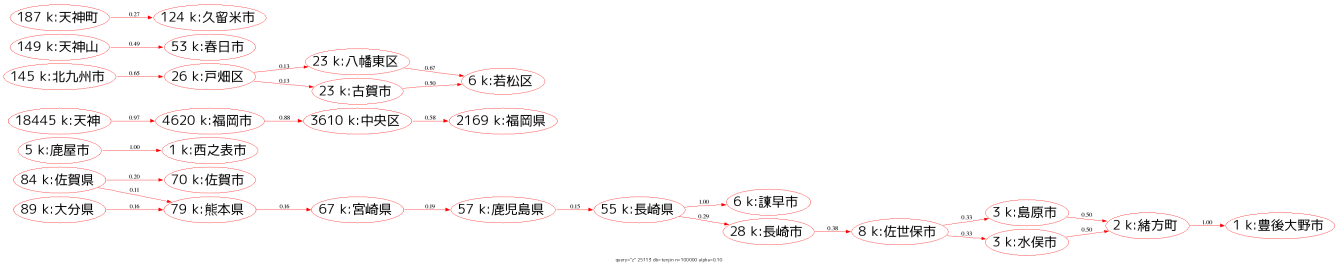

Figure 1. Classification Network of "TENJIN"

Table IV
CONTEXTS OF "TENJIN"

Subgraph Number	Place Name
1	天神町, 久留米市
	TENJIN-CHO, KURUME-CITY
2	天神山, 春日市
	TENJIN-YAMA, KASUGA-CITY
3	北九州市, 戸畑区
	KITAKYUSHU-CITY, TOBATA-WARD
3.1	八幡東区, 古賀市, 若松区
	YAHATA-HIGASI-WARD, KOGA-CITY, WAKAMATSU-WARD
4	天神, 福岡市, 中央区, 福岡県
	TENJIN, FUKUOKA-CITY, CHUO-WARD, FUKUOKA-PREFECTURE
5	鹿屋市, 西之表市
	KANOYA-CITY, NISHINOOMOTE-CITY
6	佐賀県, 佐賀市
	SAGA-PREFECTURE, SAGA-CITY
6.1	長崎県, 諫早市
	NAGASAKI-PREFECTURE, ISAHAYA-CITY
6.2	長崎市, 佐世保市, 島原市, 水俣市
	NAGASAKI-CITY, SASEBO-CITY, SHIMABARA-CITY, MINAMATA-CITY
6.3	緒方町, 豊後大野市
	OGATA-CHO, BUNGO-OHNO-CITY

since the frequency of the words are affected with the co-occurrences. The subgraphs 3.1, 5 and 6.2 contains mixed contexts with the names in difference region in the same graph. In the subgraph of the words with high frequency, we can easily interpret the place name as location. On the other hand, if the frequencies are lower, there seems to be mixture of the context and the location. This observation suggest that the classification network will be useful if we have large size of data.

VII. CONCLUSION AND FURTHER WORK

In this paper, we reported a trial of extracting place names from blog articles. We analyze four dictionaries of place names and integrated a list of place names, and extracted those names from 7,917,385 blog articles on Kyushu area. We proposed a method to use the classification network as contexts of place name in a set of documents. The classification network constructed from the blog argicles that contain the place name "TENJIN" is analyzed, where we found that 4 subgraphs are easy to interpret the location, that we can find meaningful relations in the 3 subgraphs and that the frequent words are easy to grasp.

We think that the list of 943 place names of Kyushu area are enough to start analysis. However, we need to evaluate if the restriction of the verbs to "go" and "come" were appropriate. Quantitative evaluation is needed for the effect of the classification network to disambiguate.

REFERENCES

[1] T. Baba, L. Liu, and S. Hirokawa, Formal Concept Analysis of Medical Incident Reports, 14th International Conference, KES 2010, Part III, Knowledge Based and Intelligent Information and Engineering Systems, LNCS 6278, pp.207–214. (2010)

[2] S. Esparcia, V. Sanchez-Anguix, E. Argente, A. Garcia-Fornes, V. Julian, Integrating Information Extraction Agents

into a Tourism Recommender System, Proc. HAIS2010, Springer LNAI 6077, pp.193–200. (2010)

[3] R. Grishman, and B. Sundheim, Message understanding conference-6: A brief history, Proceedings of the 16th conference on Computational linguistics-Volume 1, pp.466–471. (1996)

[4] Q. Hao, R. Cai, Ch.Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang, Equip Tourist with Knowledge Mined from Travelogues, Proc. WWW2010, pp.401–410. (2010)

[5] S. Hirokawa, T. Baba, and T. Nakatoh, Search and Analysis of Bankruptcy Cause by Classification Network, Proc. 1st International Conference on Model & Data Engineering MEDI2011, LNCS 6918, pp.152–161. (2011)

[6] A. Ishino, H. Nanba, and T. Takezawa, Automatic Compilation of Travel Information from Automatically Identified Travel Blog Entries, Journal of Japan Society for Fuzzy Theory and Intelligent Informatics 22(6), pp.667–679, (2010) (in Japanese)

[7] I. Kinjo and A. Ohuchi, Web data analysis for Hokkaido tourism information, IEICE technical report. Data engineering 101(193), pp.99–104. (2001) (in Japanese)

[8] R. Lee, H. Shiina, H. Takakura, Y. Kambayashi, Two-Dimensional Range Query Processing for Geographic Web Search, IPSJ SIG-DB 71,pp.413–420, (2003) (in Japanese)

[9] H. Okumura, M. Tokuhisa, J. Murakami, and M. Murata, Trial of extracting and classifying strong points in sightseeing area, IEICE technical report. Natural language understanding and models of communication, 110(245), pp.25–30. (2010) (in Japanese)

[10] H. Ozaku, M. Utiyama, H. Isahara, Y. Kono, and M. Kidode, An Event Information Retrieval Method Using Features of Keyword Appearance in Newspaper Corpora, Trans. of the Japanese Society for Artificial Intelligence 19, pp.225–233. (2004) (in Japanese)

[11] J. M. Ruiz-Martinez, D. Castellanos-Nieves, R. Valencia-Garcia, J. T. Fernandez-Brieis, F. Garcia-Sanchez, P. J. Vivancos- Vincente, J. S. Castejon-Garrido, J. B. Camon, R. Martinez .Bejar, Accessing Touristic Knowledge Bases through a Natural Language Interface, Proc. PKAW2008, Springer LNAI 5465, pp.147–160. (2009)

[12] H. Saito and A. Ohuchi, A Study of Visualizing Method of WWW Documents to Construct the Concept on Sightseeing Information, IPSJ SIG Notes 2001(70), pp.429-435. (2001) (in Japanese)

[13] X. Wu, S. Hirokawa, C. Yin, T. Nakatoh, and Y. Tabata, Extraction and Comparison of Tourism Information on the Web, Proc. AROB16 (16th International Symposium on Artificial Life and Robotics), pp.228–231. (2011)

[14] C. Yin, X. Wu, S. Hirokawa, and T. Nakatoh, A Proposal of "TOIEBA" Search Engine for Tourism Event, IEICE technical report. Artificial intelligence and knowledge-based processing 110(301), pp.43–47. (2010) (in Japanese)

[15] C. Yin, T. Nakatoh, S. Hirokawa, X. Wu, and J. Zeng, A proposal of search engine XYZ for tourism events, Proc. JCAI (International Joint Conference on Artificial Intelligence) Vol.1, pp.178–181. (2010)