Evaluation of tourism resources extraction based on Japanese dependency analysis

Nakatoh, Tetsuya
Research Institute for Information Technology, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings - 2nd IIAI International Conference on Advanced Applied Informatics, IIAI-AAI 2013, pp.100-103, 2013-12-16
バージョン：
権利関係：

Evaluation of Tourism Resources Extraction based on Japanese Dependency Analysis

Tetsuya Nakatoh and Sachio Hirokawa
Research Institute for Information Technology,
Kyushu University.
Email: {nakatoh, hirokawa}@cc.kyushu-u.ac.jp

***Abstract*—Blog articles by tourists contain interesting and personal experiences of where and how they have gone, what they have done and what they thought. Such individual experiences are helpful in many cases compared to the general and official information about the tourist resort by tourist agents. However, it is not easy to choose related articles and to extract still more nearly required information from these unsorted blog articles. We have proposed a method of feature extraction by dependency analysis of those sentences that describe tourist's behavior. This paper apply the proposed method to 7,917,385 blog articles on Kyushu area and shows the evaluation about the obtained resources for tourism.**

I. Introduction

The tourist resorts and the touristic institutions have special feature on their own. However, it is not easy to give a potential visitor that information. The tourist agencies already have been making advertisements and campaigns using media in order to get many tourists to come, before the Internet age. Now, they are making an effort to let many people know the special feature of the area or a store using the sightseeing portal site and the original Web page. Some tourists choose a destination using these pieces of Web information, and are enjoying the special feature of the visited area.

On the other hand, blog articles have the information about different individual experience from the information distributed officially. The tourist who actually went to the tourist resort may find by himself the special feature seldom known in addition to the known special feature got in advance, and may enjoy it. Or by a visitor's viewpoint, they may have the information and evaluation which are different from the offer side about the existing special feature. Such the special feature and information that were experienced by the tourist are based on personal experience, and were only shared only among very familiar persons before. However, generally they came to be exhibited by the blog article which described an individual travel record and individual experience now. Such information found out by the tourists is useful also for the tourist that the destination will be decided, the tourist agent who wants to improve the type of service and also for the self-governing body which wants to find out the new special feature of tourism.

However, it is not easy to extract the information appropriately from the blog articles indicated without being arranged, and to use it. We paid our attention to description of a tourist's tourism action in a blog report. Specifically, the analysis unit of tourism behavior was a three-piece set of the target (noun) of the behavior, the verb of behavior, and the particle that connects them. Their set were obtained by the dependency analysis. By extracting such basic data and analyzing them by a statistical method, the typical noun can be extracted with typical tourism behavior. In this paper, we evaluate the object noun of the tourism behavior acquired with the proposal method, and show that acquisition of useful information is possible.

II. Related Work

We can find tourism information on Web in (a) tourism portal sites, in (b) general web pages, and in (c) blog articles. There are several systems and researches intended for each target.

Esparcia et al. [2] proposed a recommendation and a clustering system, and showed their effectiveness for tourism portals. Ruiz-Martinez et al. [10] developed a natural language interface for tourism search engine. Saito and Ohuchi [11] proposed "keymaps" that visualizes co-occurrences of keywords in tourism documents. Kinjo and Ohuchi [6] analyzed the patterns in HTML documents that characterize the occurrences of NEs(Named Entity), such as the name of the location and the name of the tourism events. Hao et al. [3] and Ozaku et al. [9] studied the clue words that can be used to extract tourism related to NEs. Ishino et al. [5] reported the characteristic keywords that distinguish tourism blogs from other general blogs. Okumura et al. [8] proposed the method to extract and classify strong points in sightseeing area as support techniques to develop sightseeing area. Wu et al. [14] reported the difference between tourism information which a local government offers, and tourism information written in blog articles. Hirokawa et al. [4] proposed a search engine that focuses on the usage of onomatopoeic words that appear on tourism blogs. Yin et al. [16], [15] proposed the method of searching a characteristic tourism event in each area. The purpose of this paper is extraction of concrete behavior of tourism, and differs from these studies.

Aizawa and Nakawatase [1] tried the automatic extraction of synonyms with sample phrases using dependency analysis of text. Although this paper also uses a dependency analysis for information extraction, collection of synonyms is not the purpose.

III. Extraction Method of Tourist Behavior

A. Definition of Tourist Behavior

In many cases, tourists will do behavior similar at a tourist resort. We thought it possible to get to know behavior of tourism which is not contained in the existing tourism information by gathering such behavior. Tourism behavior is usually the combination of general action and its object. Therefore, we decided to describe tourism behavior by the set of an object and behavior.

We extract the noun n_i with Verb v_k and Particle p_j first. We defined the set (n_i, p_j, v_k) as behavior.

The behavior which appears in high frequency at the blog article about one area is surely behavior peculiar to the area. We call it tourism behavior.

B. Our Previous Study

We considered evaluating the characteristic tourism behavior for every area. Our previous study [12], [13] evaluated the noun obtained as an object of behavior after fixing the particle and verb of tourism behavior.

[12] used evaluation by the number of the tourism behavior which appears for every area. However, many general nouns which are not related to tourism were extracted.

Therefore, We improved the following two points in [13]. The first point is using the deviation of the frequency of appearance of the noun for every area for evaluation. The second point is not having used tourism behavior for evaluation of a noun, but having used the frequency of appearance of the noun in the blog article of an area. Thereby, more information can be used for selection of objects.

IV. Experiment and Evaluation

A. Basic Data

7,917,385 blog articles relevant to the Kyushu area were gathered using a Web crawler. The blog articles containing each prefecture name of Kyushu were extracted from them, and they were made into the information about each prefecture. The number of the blog articles about each prefecture is shown in Table I.

TABLE I
Number of Blog Articles

Area (Prefecture)	Number of Blog Articles
Fukuoka	152,421
Saga	22,042
Kumamoto	38,799
Ooita	76,580
Nagasaki	33,711
Miyazaki	38,954
Kagoshima	36,230
Okinawa	114,132

B. Extraction

In this section, we pay our attention to what is eaten regionally as an example of the tourism information on the area. The general nouns depending on the act "to eat" are extracted from blog articles using the method of [13]. Here, the nouns whose frequency of the occurrence is 1 are removed. It is for eliminating the word which appeared in the blog article accidentally regardless of the information on the area. The valuation method of [13] gave the rank to the obtained nouns.

The candidate for the words of tourism resources were extracted from blog articles of Kyushu area which showed by Sec. IV-A. 100 tops of 718 words obtained by the extraction are shown in table II. Most obtained nouns seem to be the dish often eaten in the areas, such as local culinary specialties. We evaluate them in the next subsection.

C. Evaluation

The accuracy as a result of the front section is evaluated by the following procedure.

1) 100 obtained pairs of nouns and area are sorted at random, and a list is made.
2) 13 examiners shown this list judge whether they are suitable respectively as what is eaten in the travel of that area.
3) Examiners have to choose from the following three items.
   - OK : It is suitable as information about tourism.
   - NG : It is unsuitable as information about tourism.
   - Unknown : I do not know it. I cannot judge it.
4) The number which chose Unknown is excepted and evaluation of each word is defined by the majority of OK and NG.

The obtained evaluation is already published to Table II. The words with evaluation of NG was specified in the column of propriety.

D. Consideration

The accuracy about extraction of 100 nouns was 86%. Probably, it will be useful enough.

We check below each 14 noun judged to be NG in detail.

Although four expressions, a "沖縄限定品 (Okinawa limited article)", the "長崎名物 (Specialty of Nagasaki)", the "鹿児島名物 (Specialty of Kagoshima)", and the "福岡名物 (Specialty of Fukuoka)", are the Japanese expressions right as an object to eat, there is no information as what is eaten by a sightseeing tour. On the other hand, it is not difficult to eliminate expression of such a fixed form.

The name of a place of "熊本 (Kumamoto)", "宮崎 (Miyazaki)", and "大分県 (Oita Prefecture)" is not an object to eat. These are considered to have been extracted by failure of dependency analysis. Moreover, euphemistic expression like "eating Kumamoto" may also be included. Exclusion is easy although each of these is the name of a places therefore.

"半額麺" is related with the coupon in which ramen noodles become half the price. The blog article which made subject

TABLE II
TARGET TO EAT

Rank	Deviation Score	TF	Noun	Area	Examiners' Judgement				
					OK	NG	Unknown	Rate of OK	Propriety
1	59.354	37	中身汁	Okinawa	5	0	8	1.00	
1	59.354	28	牛汁	Okinawa	7	0	6	1.00	
1	59.354	26	海ブドウ	Okinawa	10	0	3	1.00	
1	59.354	13	山原そば	Okinawa	7	0	6	1.00	
1	59.354	7	牛さん豚さん	Miyazaki	4	6	3	0.40	NG
1	59.354	7	本場長崎ちゃんぽん	Nagasaki	13	0	0	1.00	
1	59.354	7	ヤギ料理	Okinawa	7	1	5	0.88	
1	59.354	7	沖縄ランチ	Okinawa	7	4	2	0.64	
1	59.354	7	沖縄限定品	Okinawa	3	7	3	0.30	NG
1	59.354	5	熊本産馬刺し	Kumamoto	10	1	2	0.91	
1	59.354	5	手作り沖縄そば	Okinawa	12	0	1	1.00	
1	59.354	5	木灰そば	Okinawa	7	0	6	1.00	
1	59.354	4	半額麺	Fukuoka	5	8	0	0.38	NG
1	59.354	4	サイミン	Okinawa	3	0	10	1.00	
1	59.354	4	与那国ソバ	Okinawa	12	0	1	1.00	
1	59.354	4	名物シシリアンライス	Saga	6	1	6	0.86	
1	59.354	3	宮崎産牛	Miyazaki	10	1	2	0.91	
1	59.354	2	ねぎ丼	Fukuoka	7	1	5	0.88	
1	59.354	2	南米風レッドカレー	Fukuoka	4	2	7	0.67	
1	59.354	2	焼サバ寿司	Kumamoto	8	1	4	0.89	
1	59.354	2	年越し魚	Kumamoto	5	2	6	0.71	
1	59.354	2	宮崎マンゴードロップ	Miyazaki	10	0	3	1.00	
1	59.354	2	宮崎黒毛和牛	Miyazaki	13	0	0	1.00	
1	59.354	2	タコせんべい	Okinawa	5	0	8	1.00	
1	59.354	2	モズクてんぷら	Okinawa	8	1	4	0.89	
1	59.354	2	沖縄ちゃんぽん丼	Okinawa	9	1	3	0.90	
1	59.354	2	郷土料理屋さん	Okinawa	2	8	3	0.20	NG
1	59.354	2	白味噌ラーメン	Okinawa	7	0	6	1.00	
1	59.354	2	あなご重	Ooita	8	0	5	1.00	
1	59.354	2	氷宇治金時	Ooita	4	3	6	0.57	
1	59.354	2	卵掛けご飯	Ooita	6	3	4	0.67	
32	59.354	29	卓袱料理	Nagasaki	8	0	5	1.00	
33	59.354	627	長崎ちゃんぽん	Nagasaki	12	1	0	0.92	
34	59.353	105	長崎チャンポン	Nagasaki	11	1	1	0.92	
35	59.353	144	長崎名物	Nagasaki	5	6	2	0.45	NG
36	59.353	2250	沖縄そば	Okinawa	12	0	1	1.00	
37	59.353	218	宮崎地鶏	Miyazaki	12	0	1	1.00	
38	59.352	410	佐賀牛	Saga	12	1	0	0.92	
39	59.352	168	馬肉	Kumamoto	10	1	2	0.91	
40	59.352	95	宮崎マンゴー	Miyazaki	10	0	3	1.00	
41	59.352	115	長崎皿うどん	Nagasaki	9	0	4	1.00	
42	59.352	1804	沖縄料理	Okinawa	8	3	2	0.73	
43	59.351	599	熊本ラーメン	Kumamoto	11	1	1	0.92	
44	59.350	114	鹿児島ラーメン	Kagoshima	12	0	1	1.00	
45	59.350	22	鹿児島名産	Kagoshima	5	6	2	0.45	NG
46	59.350	77	佐賀ラーメン	Saga	9	1	3	0.90	
47	59.349	743	カステラ	Nagasaki	11	0	2	1.00	
48	59.348	42	武者返し	Kumamoto	4	3	6	0.57	
49	59.347	25	鹿児島料理	Kagoshima	7	6	0	0.54	
50	59.346	491	馬刺	Kumamoto	10	1	2	0.91	
51	59.345	413	馬刺し	Kumamoto	11	0	2	1.00	
52	59.345	628	宮崎牛	Miyazaki	12	1	0	0.92	
53	59.343	28	桃カステラ	Nagasaki	9	0	4	1.00	
54	59.339	766	黒豚	Kagoshima	11	0	2	1.00	
55	59.339	82	しろくま	Kagoshima	12	1	0	0.92	
56	59.339	68585	熊本	Kumamoto	2	11	0	0.15	NG
57	59.338	120	白熊	Kagoshima	9	3	1	0.75	
58	59.338	64827	宮崎	Miyazaki	0	13	0	0.00	NG
59	59.337	45	塩せんべい	Okinawa	6	0	7	1.00	
60	59.337	70426	熊	Kumamoto	0	8	5	0.00	NG
61	59.334	59	福岡名物	Fukuoka	5	6	2	0.45	NG
62	59.330	452	皿うどん	Nagasaki	11	1	1	0.92	
63	59.330	121	鹿児島黒豚	Kagoshima	11	1	1	0.92	
64	59.330	291	肉巻き	Miyazaki	12	0	1	1.00	
65	59.330	439	とり天	Ooita	10	0	3	1.00	
66	59.328	59	太平燕	Kumamoto	8	0	5	1.00	
67	59.324	19866	大分県	Ooita	1	11	1	0.08	NG
68	59.318	17	宮崎牛ステーキ	Miyazaki	11	0	2	1.00	
69	59.317	187	肉巻きおにぎり	Miyazaki	12	0	1	1.00	
70	59.314	45	辛麺	Miyazaki	6	2	5	0.75	
71	59.312	6	カルチャー焼き	Saga	8	0	5	1.00	
72	59.305	349	タコライス	Okinawa	4	1	8	0.80	
73	59.303	1512	ゴーヤ	Okinawa	7	1	5	0.88	
74	59.298	84	ヤンバルクイナ	Okinawa	2	6	5	0.25	NG
75	59.298	517	パイン	Okinawa	9	0	4	1.00	
76	59.290	47	赤牛	Kumamoto	6	1	6	0.86	
77	59.280	104	シシリアンライス	Saga	8	0	5	1.00	
78	59.278	42	五島うどん	Nagasaki	11	0	2	1.00	
79	59.266	17	チキン南蛮カレー	Miyazaki	11	0	2	1.00	
80	59.264	395	チキン南蛮	Miyazaki	10	0	3	1.00	
81	59.263	13	和牛ハンバーグ	Saga	9	3	1	0.75	
82	59.254	620	明太子	Fukuoka	13	0	0	1.00	
83	59.247	110	沖縄ソバ	Okinawa	12	0	1	1.00	
84	59.245	38	シロクマ	Kagoshima	11	2	0	0.85	
85	59.240	153	冷麺	Ooita	8	2	3	0.80	
86	59.232	14	クロメたこ焼き	Saga	7	1	5	0.88	
87	59.225	2040	中華	Nagasaki	7	5	1	0.58	
88	59.217	240	パパイヤ	Okinawa	8	0	5	1.00	
89	59.204	11	小浜ちゃんぽん	Nagasaki	10	0	3	1.00	
90	59.187	406	ぜんざい	Okinawa	1	5	7	0.17	NG
91	59.183	436	スイカ	Kumamoto	11	0	2	1.00	
92	59.178	10	韓国冷麺	Ooita	3	7	3	0.30	NG
93	59.176	2	リンガーハットチャンポン	Nagasaki	6	3	4	0.67	
93	59.176	2	島原名物具雑煮	Nagasaki	9	0	4	1.00	
95	59.176	434	もずく	Okinawa	5	2	6	0.71	
96	59.169	42	トウガラシ	Kagoshima	4	1	8	0.80	
97	59.159	1659	ちゃんぽん	Nagasaki	10	2	1	0.83	
98	59.155	373	納豆	Kumamoto	6	4	3	0.60	
99	59.141	74	鶏飯	Kagoshima	13	0	0	1.00	
100	59.138	149	イワシ	Nagasaki	6	0	7	1.00	

the coupon of the appendix in the informational magazine of the area has expression by ”I ate a half-the-price noodle.” It will also be possible to give a tourism meaning by adding pertinent information to this.

There were many blogs which wrote that “牛さん豚さん” (Mr. cow and Ms. piggy) of Miyazaki Prefecture were slaughtered owing to prevalence of foot and mouth disease.

The “熊” has appeared in large numbers as a part of name of a place Kumamoto. Since there was one expression by ”Someone eats a bear”, it was extracted. “ヤンバルクイナ” is a bird of rare species which inhabits only Okinawa. It was extracted from the sentence ”A cat will eat an Okinawa rail.” These are the examples extracted from one occurrence. Although it has agreed for the purpose of this method of extraction of niche information, in order to eliminate such an unnecessary object, a certain examination is required.

“郷土料理屋さん (The local-culinary-specialties restaurant)” had appeared as expression of “trying the local-culinary-specialties restaurant at various restaurants.” That is, it was not an object of ”eating.” Adjustment of a dependency analysis is required.

Although they were judged to be NG by the examiners, when the author investigated “ぜんざい (sweet red-bean soup with rice cakes)” and the “韓国冷麺 (South Korean Korean-style buckwheat noodles), they were suitable as resources for tourism respectively. “ぜんざい (The sweet red-bean soup with rice cakes)” of Okinawa is popular also in tourism by a kind of the oyster ice the “沖縄ぜんざい.” “韓国冷麺 (The South Korean Korean-style buckwheat noodles)” of Oita is the “別府冷麺 (Beppu Korean-style buckwheat noodles)” of Beppu-shi, Oita. It is popular as a local gourmet too. Such extraction of the food seldom known is very significant.

V. Conclusion and Future Works

We have been working on content extraction of the tourism information from the blog which used a Japanese dependency analysis and the deviation of the occurrence. In this paper, evaluation by 13 examiners showed that the accuracy of the top 100 extraction result using the proposal method was 86%.

On the other hand, evaluation in this paper was limited to the extraction result of having made one kind of a particle and a verb pair into the key. You have to examine the extraction result by more keys. Moreover, the method of searching for the key which extracts tourism information is also required. We are planning generalization of the resources-for-tourism extraction method from blog articles based on a tourist’s behavior further.

References

[1] A. Aizawa and H. Nakawatase, “Automatic extraction of synonyms with sample phrases using dependency analysis of text and its application to large-scale corpora,” in *the 20th Annual Conference of the Japanese Society for Artificial Intelligence*, 2E1-5, 2006. (in Japanese)

[2] S. Esparcia, V. Sanchez-Anguix, E. Argente, A. Garcia-Fornes and V. Julian, “Integrating Information Extraction Agents into a Tourism Recommender System,” in *Proc. HAIS2010, Springer LNAI 6077*, pp.193-200, 2010.

[3] Q. Hao, R. Cai, Ch. Wang, R. Xiao, J.-M. Yang, Y. Pang and L. Zhang, “Equip Tourist with Knowledge Mined from Travelogues,” in *Proc. WWW2010*, pp.401-410, 2010.

[4] S. Hirokawa, C. Yin, K. Hashimoto and K. Takeuchi, “Search and Analysis of Gourmet Blogs with a Particular Reference to Onomatopoeia,” *ICIC Express Letters, Volume 5, Issue 8(B)*, pp.2971-2976, 2011.

[5] A. Ishino, H. Nanba, H. Gaguma, T. Ozaki, D. Kobayashi and T. Takezawa, “Automatic Compilation of Travel Information from Automatically Identified Travel Blogs,” *IEICE Tech Report, WI2-2009*, pp.19-23, 2009.(in Japanese)

[6] I. Kinjo and A. Ohuchi, “Web data analysis for Hokkaido tourism information,” *IEICE Tech. Report, DE2001-07*, pp.99-104, 2001. (in Japanese)

[7] T. Kudo and Y. Matsumoto, “Fast Methods for Kernel-Based Text Analysis,” Proceedings of *the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp.24-31, 2003.

[8] 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)

[9] H. Ozaku, M. Utiyama and M. Kidode, “An Event Information Retrieval Method Using Features of Keyword Appearance in Newspaper Corpora,” *Trans. JSAI, AI19*, pp.225-233, 2004. (in Japanese)

[10] 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 and R. Martinez-Bejar, “Accessing Touristic Knowledge Bases through a Natural Language Interface,” *Springer LNAI 5465*, pp.147-160, 2009.

[11] H. Saito and A. Ohuchi, “A Study of Visualizing Method of WWW Documents to Construct the Concept on Sightseeing Information,” *IEICE Tech. Report, DE2001-07*, pp.261-267, 2001. (in Japanese)

[12] T. Nakatoh, C. Yin and S. Hirokawa, “Characteristic Grammatical Context of Tourism Information,” *ICIC Express Letters*, Vol.6, No.3. March 2012, pp.753-758, 2012.

[13] T. Nakatoh and S. Hirokawa, “Extraction of Tourist Behavior Contexts from Blog by Verbs and Their Objects,” Proc. of *IIAI International Conference on Advanced Applied Informatics*, September 2012.

[14] X. Wu, S. Hirokawa, C. Yin, T. Nakatoh and Y. Tabata, “Extraction and Comparison of Tourism Information on the Web,” in *Proc. of AROB2011*, 2011.

[15] C. Yin, T. Nakatoh, S. Hirokawa, X. Wu and J. Zeng, “A proposal of search engine “ XYZ ” for tourism events,” *Second IITA International Joint Conference on Artificial Intelligence*, 2010.

[16] 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* vol. 110, no. 301, pp.43-47, 2010. (in Japanese)