Chapter 6

Automatic Compilation of Travel Information from Texts: A Survey

Hidetsugu Nanba, Aya Ishino and
Toshiyuki Takezawa

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51290

1. Introduction

Travel guidebooks and portal sites provided by tour companies and governmental tourist boards are useful sources of information about travel. However, it is costly and time-consuming to compile travel information for all tourist spots and to keep these data up-to-date manually. Recently, research about services for the automatic compilation and recommendation of travel information has been increasing in various research communities, such as natural language processing, image processing, Web mining, geographic information systems (GISs), and human interfaces. In this chapter, we overview the state of the art of the research and several related services in this field. We especially focus on research in natural language processing, including text mining.

The remainder of this chapter is organized as follows. Section 2 explains the automatic construction of databases for travel. Section 3 describes analysis of travelers' behavior. Section 4 introduces several studies about recommending travel information. Section 5 shows interfaces for travel information access. Section 6 lists several linguistic resources. Finally, we provide our conclusions and offer future directions in Section 7.

2. Automatic construction of databases for travel

In this section, we describe several studies about constructing databases for travel. In Section 2.1, we introduce a study that identified travel blog entries in a blog database. In Section 2.2, we describe several methods to construct databases for travel by extracting travel information, such as tourist spots or local products, from travel blog entries using information extraction techniques. In Section 2.3, we explain a method that constructs travel links automatically.

2.1. Automatic identification of travel blog entries

Travel blogs[1] are defined as travel journals written by bloggers in diary form. Travel blogs are considered useful for obtaining travel information, because many bloggers' travel experiences are written in this form.

There are various portal sites for travel blogs, which we will describe in Section 6. At these sites, travel blogs are manually registered by bloggers themselves, and the blogs are classified according to travel destination. However, there are many more travel blogs in the blogosphere, beyond these portal sites. In an attempt to construct an exhaustive database of travel blogs, Nanba et al. [25] identified travel blog entries written in Japanese in a blog database.[2]

Blog entries that contain cue phrases, such as "travel", "sightseeing", or "tour", have a high degree of probability of being travel blogs. However, not every travel blog contains such cue phrases. For example, if a blogger describes his/her journey to Norway in multiple blog entries, the blog might state "We traveled to Norway" in the first entry, while only writing "We ate wild sheep!" in the second entry. In this case, because the second entry does not contain any expressions related to travel, it is difficult to identify it as a travel blog entry. Therefore, Nanba et al. focused not only on each blog entry but also on the surrounding entries for the identification of travel blog entries. They formulated the identification of travel blog entries as a sequence-labeling problem, and solved it using machine learning. For the machine learning method, they examined the Conditional Random Fields (CRF) method [20]; its empirical success has been reported recently in the field of natural language processing. The CRF-based method identifies the tag [3] of each entry. Features and tags are given in the CRF method as follows: (1) k tags occur before a target entry; (2) k features occur before a target entry; and (3) k features follow a target entry (see Figure 1). They used the value of $k = 4$, which was determined in a pilot study. Here, they used the following features for machine learning: whether an entry contains any of 416 cue phrases, such as "旅行 (travel)", "ツアー (tour)", and "出発 (departure)", and the number of location names in each entry.

Using the above method, Nanba et al. identified 17,268 travel blog entries from 1,100,000 blog entries, and constructed a system that plotted travel blog entries on a Google map (see Figure 2).[4] In this figure, travel blog entries are shown as icons. If the user clicks an icon, the corresponding blog entry is shown in a pop-up window.

2.2. Automatic extraction of travel information from texts

Nakatoh et al. [24] proposed a method for extracting names of local culinary dishes from travel blogs written in Japanese, which were identified when the blog entry included both the name of a sightseeing destination and the word "tourism". They extracted local dishes by gathering nouns that are dependent on the verb "を食べる" (eat). Tsai and Chou [32] also proposed a method for extracting dish names from restaurant review blogs written in Chinese using a machine learning (CRF) technique.

[1] We use the term *travel blog*. Other studies use the term "Travelogues" [10], indicating social networking service (SNS) content, blogs, reviews, message boards, and so on, for travel.

[2] Although Nanba et al. identified Japanese travel blogs, their method can be applied to blogs written in other languages, if cue phrases for the language are prepared.

[3] In this case, the tag indicates whether each entry is a travel blog entry or not.

[4] http://www.ls.info.hiroshima-cu.ac.jp/test/travel-map/xml-travelmap.html

http://dx.doi.org/10.5772/51290

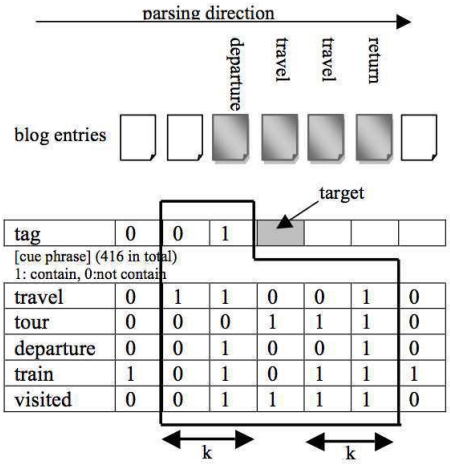

Figure 1. Features and tags used in CRF

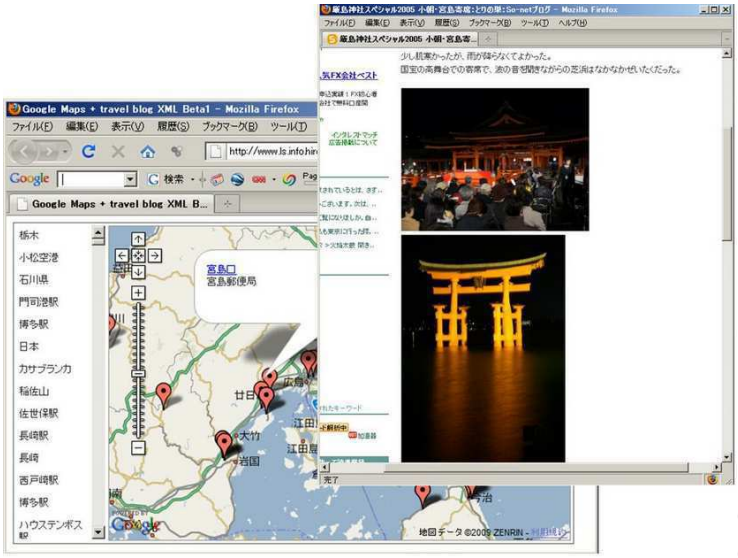

Figure 2. Travel blog entries plotted on a Google map

In the following, we explain the detail of the bootstrapping-based and machine learning-based information extraction approaches based on Nanba's work [25]. Nanba et al. extracted pairs comprising a location name and a local product from travel blogs written in Japanese, which were identified using the method described in Section 2.1. For the efficient extraction of travel information, they employed a bootstrapping method.

First, they prepared 482 pairs as seeds for the bootstrapping. These pairs were obtained automatically from a "Web Japanese N-gram" database provided by Google, Inc. The

database comprises *N*-grams (*N* = 1–7) extracted from 20 billion Japanese sentences on the Web. They applied the pattern "[地名] 名物「[名物]」" ([slot of "location name"] local product 「[slot of "local product"]」) to the database, and extracted location names and local products from each corresponding slot, thereby obtaining the 482 pairs.

Second, they applied a machine learning-based information extraction technique to the travel blogs identified in the previous step, and obtained new pairs. In this step, they prepared training data for the machine learning in the following three steps.

1. Select 200 sentences that contain both a location name and a local product from the 482 pairs. Then automatically create 200 tagged sentences, to which both "location" and "product" tags are assigned.[5]
2. Prepare another 200 sentences that contain only a location name. Then create 200 tagged sentences, to which the "location" tag is assigned.
3. Apply machine learning to the 400 tagged sentences, and obtain a system that automatically allocates "location" and "product" tags to given sentences.

As a machine learning method, they used CRF. The CRF-based method identifies the class of each word in a given sentence. Features and tags are given in the CRF method as follows: (1) k tags occur before a target word; (2) k features occur before a target word; and (3) k features follow a target word. They used the value of $k = 2$, which was determined in a pilot study. They used the following six features for machine learning.

- Word.
- The part of speech to which the word belongs (noun, verb, adjective, etc.)
- Whether the word is a quotation mark.
- Whether the word is a cue word, such as "名物", "名産", "特産" (local product), "銘菓" (famous confection), or "土産" (souvenir).
- Whether the word is a surface case.
- Whether the word is frequently used in the names of local products or souvenirs, such as "cake" or "noodle".

2.3. Automatic compilation of travel links

Collections of Web links are usefel information sources. However, maintaining these collections manually is costly. Therefore, an automatic method for compiling collections of Web links is required. In this section, we introduce a method that compiles travel links automatically.

From travel blog entries, which were automatically identified using the method mentioned in Section 2.1, Ishino et al. [15] extracted the hyperlinks to useful Web sites for a tourist spot included by bloggers, and thereby constructed collections of hyperlinks for tourist spots. The procedure for classifying links in travel blog entries is as follows.

[5] Here, a location name corresponds to only a local product in each sentence.

http://dx.doi.org/10.5772/51290

1. Input a travel blog entry.
2. Extract a hyperlink and any surrounding sentences that mention the link (a citing area).
3. Classify the link by taking account of the information in the citing area.

They classified link types into the following four categories.

- S (Spot): The information is about tourist spots.
- H (Hotel): The information is about accommodation.
- R (Restaurant): The information is about restaurants.
- O (Other): Other than types S, H, and R.

A hyperlink may be classified as more than one type. For example, a hyperlink to "ラーメン博物館" (Chinese noodle museum, http://www.raumen.co.jp/home/) was classified as types S and R, because the visitors to this museum can learn the history of Chinese noodles in addition to eating them.

For the classification of link types, they employed a machine learning technique using the following features.

- A word.
- Whether the word is a cue phrase, detailed as follows, where the numbers in brackets shown for each feature represent the number of cues.

Cue phrase	The number of cues
A list of tourist spots, collected from Wikipedia.	17,371
Words frequently used in the name of tourist spots, such as "動物園" (zoo) or "博物館" (museum).	138
Words related to sightseeing, such as "見学" (sightseeing) or "散策" (stroll).	172
Other words.	131

Table 1. Cues for type S

Cue phrase	The number of cues
Words that are frequently used in the name of hotels, such as "ホテル" (hotel) or "旅館" (Japanese inn).	9
Component words for accommodations, such as "フロント" (front desk) or "客室" (guest room).	29
Words that are frequently used when tourists stay in accommodation, such as "泊る" (stay) or "チェックイン" (check in).	14
Other words.	21

Table 2. Cues for type H

Based on this method, Ishino et al. constructed a travel link search system.[6] The system generated a list of URLs for Web sites related to a location, and automatically identified link types and the context of citations ("citing areas"), where the blog authors described the sites. Figure 3 shows a list of links related to "大阪" (Osaka).

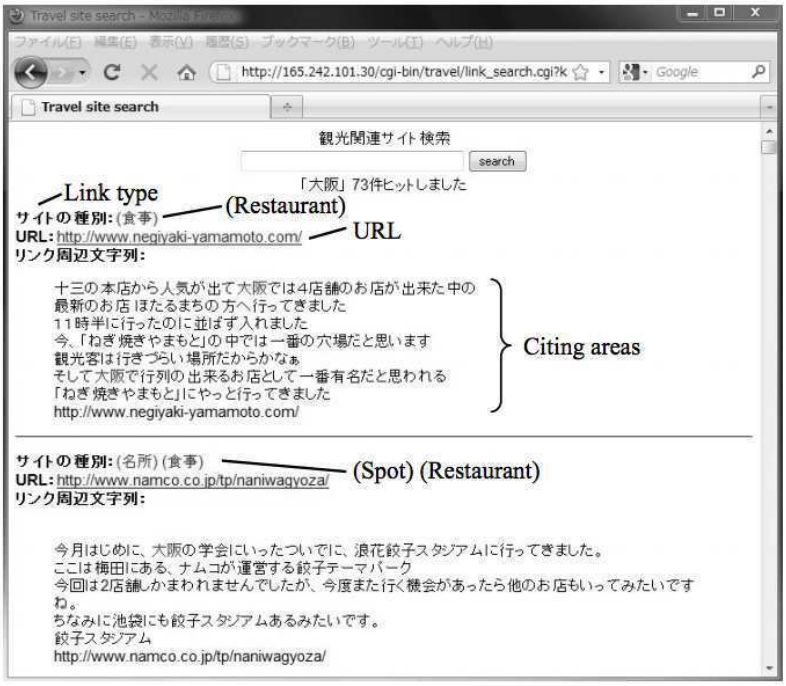

Figure 3. A list of Web sites for a travel spot

Cue phrase	The number of cues
Dish names such as "omelet", collected from Wikipedia.	2,779
Cooking styles such as "Italian cuisine", collected from Wikipedia.	114
Words that are frequently used in the name of restaurants, such as "レストラン" (restaurant) or "食堂" (dining room).	21
Words that are used when taking meals, such as "食べる" (eat) or "おいしい" (delicious).	52
General words that indicate food, such as "ご飯" (rice) or "料理" (cooking).	31
Other words.	31

Table 3. Cues for type R

[6] http://www.ls.info.hiroshima-cu.ac.jp/travel/

http://dx.doi.org/10.5772/51290

3. Travelers' behavior analysis

The analysis of people's transportation information is considered an important issue in various fields, such as city planning, architectural planning, car navigation, sightseeing administration, crime prevention, and tracing the spread of infection of epidemics. In this section, we focus on the analysis of travelers' behavior.

Ishino et al. [15] proposed a method to extract people's transportation information from automatically identified travel blogs written in Japanese [25]. They used machine learning to extract information, such as "departure place", "destination", or "transportation device", from travel blog entries. First, the tags used in their examination are defined.

- FROM tag indicates the departure place.
- TO tag indicates the destination.
- VIA tag indicates the route.
- METHOD tag indicates the transportation device.
- TIME tag indicates the time of transportation.

The following is a tagged example.

> It took <TIME>five hours< /TIME> to travel from <FROM>Hiroshima< /FROM> to<TO>Osaka< /TO> by <METHOD>bus< /METHOD>.

They formulated the task of identifying the class of each word in a given sentence and solved it using machine learning. For the machine learning method, they used CRF [20], in the same way as Nanba et al. [25], which we mentioned in Section 2.2. The CRF-based method identifies the class of each entry. Features and tags are used in the CRF method as follows: (1) k tags occur before a target entry; (2) k features occur before a target entry; and (3) k features follow a target entry. They used the value $k = 4$ [7] , which was determined via a pilot study. They used the following features for machine learning.

- A word.
- The part of speech to which the word belongs (noun, verb, adjective, etc.).
- Whether the word is a quotation mark.
- Whether the word is a cue phrase.

The details of cue phrases, together with the number of cue phrases of the given type, are shown as follows.

1. **FROM**: The word is a cue that often appears immediately after the "FROM" tag, such as "から" (from) or "を出 発" (left): 40.

[7] Nanba et al.[25] used the smaller value $k = 2$ in the extraction of pairs comprising a location name and a local product (Section 2.2), because the tags are determined by a word itself or its adjacent words in most cases in Nanba's task.

2. **FROM & TO**: The word is frequently used in the name of a tourist spot, such as “博物館” (museum) or “遊園地” (amusement park): 45.
   The word is frequently used in the name of a destination, such as “観光” (sightseeing tour) or “駅” (station): 11.
   The word is the name of a tourist spot: 13,779.
   The word is the name of a station or airport: 9437.
3. **TO**: The word is a cue that often appears immediately after the “TO” tag, such as “まで” (to) or “に到着” (arrival): 271.
4. **VIA**: The word is a cue that often appears immediately after the “via” tag, such as “経由” (via) or “通って” (through): 43.
   The word is the name of a highway: 101.
5. **METHOD**: The word is the name of a transportation device, such as “飛行機” (airplane) or “自動車” (car): 148.
   The word is the name of a vehicle: 128.
   The word is the name of a train or bus: 2033.
6. (TIME): The word is an expression related to time, such as “分” (minute) or “時間” (hour): 77.

They also constructed a visualization of transportation information, which is shown in Figure 4. In this figure, each arrow indicates a link from a departure place to a destination. In addition to arrows, transportation methods, such as trains or buses, are shown as icons.

Transportation information can also be extracted from texts written in English. Davidov [6] presented an algorithm framework that enables automated acquisition of map-link information from the Web, based on linguistic patterns such as “from X to”. Given a set of locations as initial seeds, he retrieved an extended set of locations from the Web, and produced a map-link network that connected these locations using edges showing the transportation type.

4. Recommending travel information

Recommendation systems provide a promising approach to ranking commercial products or documents according to a user’s interests. In this section, we describe several studies and services that recommend travel information. We describe the recommendation of tourist spots, landmarks, travel products, accomodation, and photos.

4.1. Recommending tourist spots

Recommending tourist spots[8] has been well studied in the multimedia field. Movies and images are used as information sources in addition to texts. In this section, we describe two multimedia studies.

Hao et al. [10] proposed a method for mining location-representative knowledge from travel blogs based on a probabilistic topic model (the Location–Topic model). Using this model,

[8] Here, we use the terms “tourist spot” and “landmark” for a region, such as “Paris” or “New York”, and also for a location or building, such as “the Eiffel Tower” or “Statue of Liberty”.

http://dx.doi.org/10.5772/51290

Figure 4. Example of transportation information automatically extracted from travel blogs

they developed three modules: (1) destination recommendation for flexible queries; (2) characteristics summarization for a given destination, with representative tags and snippets; and (3) identification of informative parts of a travel blog and enriching recommendations with related images.

Figure 5 shows an example of the system output. In this figure, a travel blog segment[9] is enriched with three images that depict its most informative parts. Each image's original tags and the words in the text to which it corresponds are also presented.

Wu et al. [34] proposed a system that summarized tourism-related information. When a user (traveler) entered a query, such as "What is the historical background of Tian Tan?", the system searched for and obtained information from Wikipedia, Flickr, YouTube, and official tourism Web sites using the tourist spot name as a query. The system also classified the query as belonging to one of five categories—"general", "history", "landscape", "indoor scenery", and "outdoor scenery"—in order to provide users with more relevant information. For example, when a query is classified as belonging to the "history" category, the information is obtained from texts, while for a query regarding "outdoor scenery", the information is obtained from photos and videos.

[9] A segment of a Maui travel blog entitled "Our Maiden Journey to Magical Maui", http://www.igougo.com/journal-j23321-Maui-Our_Maiden_Journey_to_Magical_Maui.html

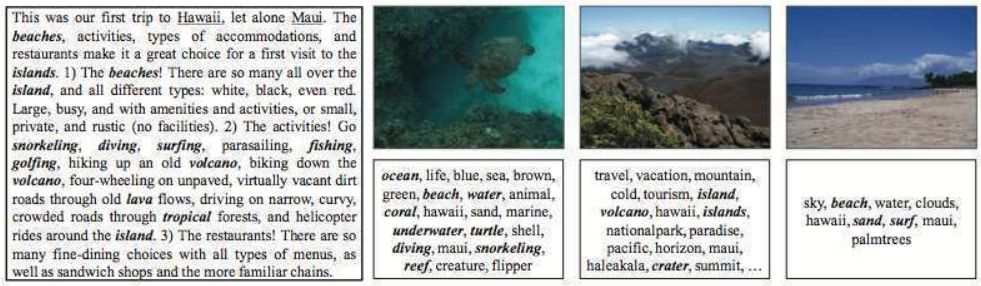

Figure 5. Example of travel blog segment visually enriched with related images

4.2. Recommending landmarks

Finding and recommending landmarks is considered an important research topic in the multimedia field, along with recommending tourist spots. Abbasi et al. [1] focused on the photo-sharing system Flickr, and proposed a method to identify landmark photos using tags and social Flickr groups. Gao et al. [7] also proposed a method to identify landmarks using Flickr and the Yahoo Travel Guide.

Ji et al. [17] proposed another method for finding landmarks. They adopted the method of clustering blog photos relating to a particular tourist site, such as Louvre Museum in Paris.[10] Then they represented these photos as a graph based on the clustering results, and detected landmarks using link analysis methods, such as the PageRank [3] and HITS [19] algorithms.

4.3. Recommending travel products

Ishino et al. [14] proposed a method that added links to advertisements for travel products to the travel information links that were described in Section 2.3.[11] The procedure for providing ad links is as follows.

1. Input a link type and the citing areas of a travel information link.
2. Extract keywords from the citing areas.
3. Extract product data containing all keywords, and calculate the similarity between the citing areas of a travel information link and the product data.
4. Provide the ad link to the product data having the highest similarity to the travel information link.

They extracted keywords for travel products corresponding to the link type. They used the same cues to classify travel information links [15] (see Section 2.3), and then extracted keywords from the citing areas of links of types S (Spot) and R (Restaurant).

[10] For calculating the similarity between two photos, they used the Bag-of-Visual-Words representation [18, 26], which represents an image as a set of salient regions (visual words), called Bag-of-Visual-Words vectors. Then the similarity between photos is measured based on the cosine distance between their Bag-of-Visual-Words vectors. In addition to the features in each image, they also used textual information for each photo, such as the title, description, and surrounding text.

[11] http://www.ls.info.hiroshima-cu.ac.jp/travel/

http://dx.doi.org/10.5772/51290

First, the method for extracting keywords from the citing areas of links of type S is described. The cues for type S, such as tourist spots collected from Wikipedia and words frequently used in the names of tourist spots, tend to become keywords. Therefore, they registered these cues as candidate keywords for links of type S. If the citing areas of these links contained candidate keywords, they extracted the candidates as keywords. In addition, if citing areas contained names of places, they extracted the names as keywords.

The cues for type R, such as dish names and cooking styles, also tend to become keywords. Therefore, they registered these cues as candidate keywords for links of type R. If the citing areas for links of type R contained candidate keywords, they extracted them as keywords.

4.4. Recommending accommodation

Titov and McDonald [31] proposed an aspect-based summarization system, and applied the method to the summarization of hotel reviews. The system took as input a set of user reviews for a specific product or service with a numeric rating (left side in Figure 6), and produced a set of relevant aspects, which they called an aspect-based summary (right side in Figure 6). To extract all relevant mentions in each review for each aspect, they introduced a topic model. They applied their method to hotel reviews on the TripAdvisor Web site[12], and obtained aspect-based summaries for each hotel.

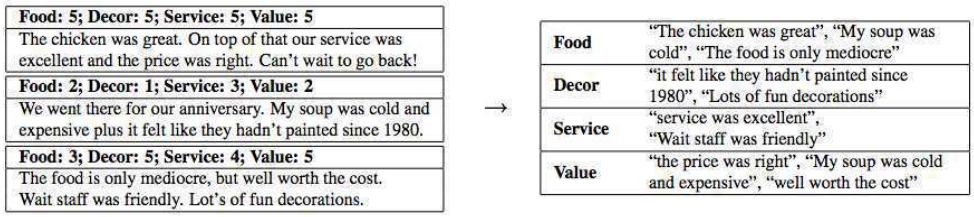

Figure 6. Producing aspect mentions from a corpus of aspect rated reviews

To obtain more reliable hotel reviews, opinion spams should be detected and eliminated. Opinion spams are fictitious opinions that have been deliberately written to sound authentic. Ott et al. [27] proposed a method to detect opinion spam among consumer reviews of hotels. They created 400 deceptive opinions using the Amazon Mechanical Turk (AMT) crowdsourcing service[13] by asking anonymous online workers (Turkers) to create the opinion spam for 20 chosen hotels. In addition to these spam messages, they selected 6,977 truthful opinions from TripAdvisor, and used both groups for their task.

4.5. Recommending photos

Bressan et al. [2] proposed a travel blog assistant system that facilitated the travel blog writing by selecting for each blog paragraph the most relevant images from an image set. The procedure is as follows.

[12] http://www.tripadvisor.com
[13] https://www.mturk.com/

1. The system adds metadata to the traveler's photos based on a generic visual categorizer, which provides annotations (short textual keywords) related to some generic visual aspects of and objects in the image.[14]

2. Textual information (tags) was obtained using a cross-content information retrieval system using a repository of multimedia objects.

3. For a given paragraph, the system ranked the uploaded images according to the similarity between the extracted metadata and the paragraph.

5. Interfaces for travel information access

In this section, we describe two studies that focused on interfaces for travel information access.

5.1. Providing travel information along streetcar lines

Ishino et al. [13] proposed a method for collecting blog entries about the Hiroshima Electric Railway (Hiroden) from a blog database.[15] Hiroden blog entries were defined as travel journals that provide regional information for streetcar stations in Hiroshima. The task of collecting Hiroden blog entries was divided into two steps: (1) collection of blog entries; and (2) identification of Hiroden blog entries.

Figure 7 shows a route map used by the system for providing travel information along the Hiroden streetcar lines. The route map shows Hiroden streetcar stations and major tourist spots. The steps in the search procedure are as follows.

- (Step 1) Click the Hiroden streetcar station, such as "原爆ドーム前" (Atomic Bomb Dome), in Figure 7 to generate a list of links to Hiroden blog entries (Figure 8).

- (Step 2) Click the link to a Hiroden blog entry to display it.

5.2. Natural language interface for accessing databases

Several ontologies for e-tourism have been developed (see Section 6). Unfortunately, the gap between human users who want to retrieve information and the Semantic Web is yet to be cloased. Ruiz-Martínez et al. [30] proposed a method for querying ontological knowledge bases using natural language sentences. For example, when the user inputted the query "I want to visit the most important tourist attractions in Paris", the system conducted part-of-speech tagging, lemmatizing, and modification of query terms by synonyms, and finally searched the ontology.

[14] Bressan et al. used images that were categorized into 44 classes as training data for visual categorization. Each class was given a short text name, such as "clouds and sky" or "beach". When an image was categorized as belonging to classes A and B using the visual categorizer, the short texts given to each class were assigned as keywords of the image.

[15] http://165.242.101.30/travel/hiroden/

http://dx.doi.org/10.5772/51290

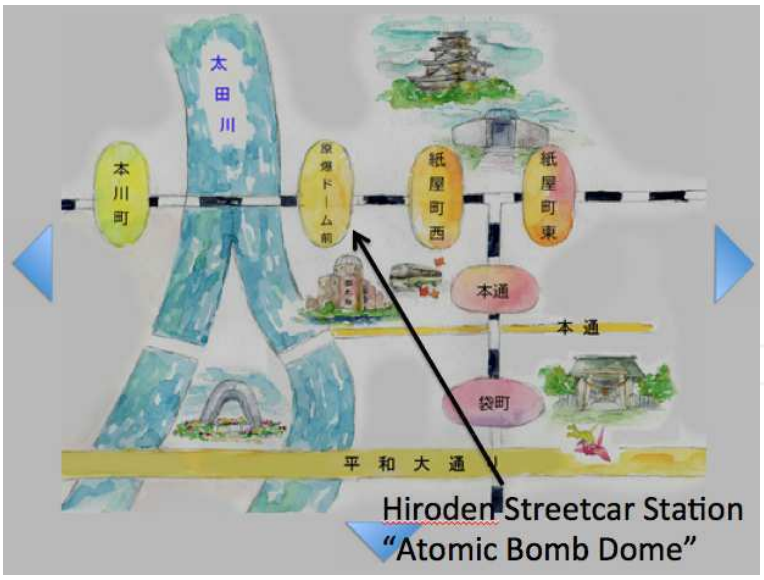

Figure 7. A route map of the Hiroden system

Figure 8. A list of links to Hiroden blog entries

6. Linguistic resources for studies of automatic compilation of travel information from texts

Text Corpora

- TripAdvisor: http://tripadvisor.com
  This site provides fifty million reviews written in various languages.
- Footstops: http://footstops.com
  This site provides more than 8,000 blog entries written in English.
- IgoUgo: http://www.igougo.com
  This site provides 530,000 reviews and 62,000 blog entries written in English.
- Travbuddy: http://www.travbuddy.com
  This site provides more than 90,000 reviews and 180,000 blog entries written in English.
- TravelBlog: http://www.travelblog.org
  This site provides more than 600,000 blog entries written in English. Each entry is classified at city level in a geographic hierarchy.
- Travellerspoint: http://www.travellerspoint.com
  This site provides more than 180,000 blog entries written in English.
- TravelPod: http://www.travelpod.com
  This site is one of the oldest travel portal, started since 1997, and provides blog entries written in English.
- 4travel: http://4travel.jp
  This site provides approximately 300,000 reviews and 600,000 blog entries written in Japanese. Each review is classified at city level in a geographic hierarchy.

Databases for Travel

- Rakuten travel data: http://www.nii.ac.jp/cscenter/idr/datalist.html (Japanese)
  Basic information about 11,468 properties and 350,000 reviews
- Travel product data in Rakuten Shopping Mall (Rakuten Ichiba): http://www.nii.ac.jp/cscenter/idr/datalist.html (Japanese)
  The data comprise 50 million items. Each item has name, code, price, URL, picture, shop code, category ID, and descriptive text and registration data.

Useful Sites or Services for Travel

- Yahoo Travel Guide: http://travel.yahoo.com/
  This site provides an area-based recommendation service. For each country, several main cities are listed.
- WikiTravel: http://wikitravel.org
  The travel recommendation system contributed by "WikiTravellers". For each destination, the articles in WikiTravel generally include all or parts of the following information: history, climate, landmarks, work information, shopping information, food, and how to get there.

http://dx.doi.org/10.5772/51290

Ontologies for Travel

- The World Tourism Organization (WTO) provides a multilingual thesaurus in English, French, and Spanish that provides a standard terminology for tourism [33].
- DERI's e-Tourism Working group has created a tourism ontology called "OnTour" [28]. This ontology describes the main conventional concepts for tourism such as accommodation or activities, together with other supplementary concepts such as GPS coordinates or a postal address.
- LA_DMS is an ontology for tourism destinations that was developed for the Destination Management System (DMS). This system adapts information requests about tourist destinations to users' needs [16].

Many other ontologies for travel were introduced by Ruiz-Martínez et al. [30].

Evaluation Workshop

GeoCLEF: Geographic Information Retrieval

GeoCLEF (http://ir.shef.ac.uk/geoclef/) was the cross-language geographic retrieval track run as part of the Cross-Language Evaluation Forum (CLEF). It operated from 2005 to 2008 [11, 12, 21, 22]. The goal of this task was to retrieve news articles relevant to particular aspects of geographic information.

NTCIR GeoTime

NTCIR GeoTime was another cross-language geographic retrieval track run as part of the NTCIR. It operated from 2008 to 2011 [8, 9]. The focus of this task was searching with geographic and temporal constraints using Japanese and English news articles as target documents.

7. Conclusions and future directions

In this chapter, we have introduced the state of the art of research and services related to travel information. There are several future directions for this research field.

- We mentioned in Section 2 that several natural language processing technologies are useful for creating databases for travel. These technologies may also be applied to maintain manually created databases or ontologies for travel, such as those discussed in Section 6.
- Multilingualization of the ontologies for travel using machine translation techniques [4] is also considered an important task for encouraging further studies in this research field.
- There are many different locations that have the same name (place name polysemy), and there may be multiple names for a given location (place name synonymy). To eliminate this geo-ambiguity problem, Ji et al. [17] proposed the Hierarchical-comparison Geo-Disambiguation (HGD) algorithm, which distinguished the city-level location using a combination of its lower-level locations, derived from the hierarchical location relationships. In addition to this method, several natural language processing

technologies, such as automatic acquisition of synonyms [5, 29, 35, 36] and word sense disambiguation [23], are available.

- Recommending landmarks (landmark finding) is a standard research topic in image processing using Flickr. In this chapter, we mentioned three studies [1, 7, 17] that relied mainly on image processing and tag-based recommendation techniques rather than natural language processing. The authors believe that there is still room to improve the methods of recommending landmarks by natural language processing, because sentiment analysis techniques, such as those used for recommending accommodation, have not yet been used for recommending landmarks.

Author details

Hidetsugu Nanba*,
Aya Ishino and Toshiyuki Takezawa

* Address all correspondence to: nanba@hiroshima-cu.ac.jp

Graduate School of Information Sciences, Hiroshima City University, Japan

References

[1] Abbasi, R., Chernov, S., Nejdl, W., Paiu, R., Staab, S. (2009) Exploiting Flickr Tags and Groups for Finding Landmark Photos. Proceedings of ECIR 2009, pp.654–661.

[2] Bressan, M., Csurka, G., Hoppenot, Y., and Renders, J.M. (2008) Travel Blog Assistant System (TBAS) - An Example Scenario of How to Enrich Text with Images and Images with Text using Online Multimedia Repositories. Proceedings of VISAPP Workshop on Metadata Mining for Image Understanding.

[3] Brin, S. and Page, L. (1998) The Anatomy of a Large-scale Hypertextual Web Search Engine. Proceedings of World Wide Web Conference 1998.

[4] Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., and Mercer, R.L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, Vol.19, No.2, pp.263–311.

[5] Callison-Burch, C., Koehn, P., and Osborne, M. (2006) Improved Statistical Machine Translation Using Paraphrases. Proceedings of NAACL 2006, pp.17–24.

[6] Davidov, D. (2009). Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.267–275.

[7] Gao, Y., Tang, J, Hong, R., Dai, Q., Chua, T.-S., and Jain, R. (2010) W2Go: A Travel Guidance System by Automatic Landmark Ranking. Proceedings of ACM Multimedia'10.

[8] Gey, F., Larson, R., Machado, J., and Yoshioka, M. (2011) NTCIR9-GeoTime Overview: Evaluating Geographic and Temporal Search: Round 2. Proceedings of NTCIR-9 Workshop Meeting.

[9] Gey, F., Larson, R., Kando, N., Machado, J., and Sakai, T. (2010) NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. Proceedings of NTCIR-8 Workshop Meeting.

[10] Hao, Q., Cai, R., Wang, C., Xiao, R., Yang, J.-M., Pang, Y., and Zhang, L. (2010) Equip Tourists with Knowledge Mined from Travelogues. Proceedings of World Wide Web Conference 2010.

[11] Gey, F., Larson, R.R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Nunzio, G.M.D., Ferro, N. (2006) GeoCLEF 2006: The CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. Proceedings of CLEF 2006, pp.852–876.

[12] Gey, F. Larson, R.R., Sanderson, M., Joho, H., Clough, P., and Petras, V. (2005) GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. Lecture Notes in Computer Science, LNCS4022, pp.908–919.

[13] Ishino, A., Nanba, H., and Takezawa, T. (2012) Construction of a System for Providing Travel Information along Hiroden Streetcar Lines. Proceedings of the 3rd IIAI International Conference on e-Services and Knowledge Management.

[14] Ishino, A., Nanba, H., and Takezawa, T. (2011) Providing Ad Links to Travel Blog Entries Based on Link Types. Proceedings of the 9th Workshop on Asian Language Resources, collocated with IJCNLP 2011, pp.63–70.

[15] Ishino, A., Nanba, H., and Takezawa, T. (2011) Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries. Proceedings of ENTER 2011.

[16] Jakkilinki, R., Ceorgievski, M., and Sharda, N. (2007) Connecting Destinations with an Ontology-Based e-Tourism Planner. Information and Communication Technologies in Tourism, pp.21–32.

[17] Ji, R., Xie, X., Yao, H., and Ma, W.-Y. (2009) Mining City Landmarks from Blogs by Graph Modeling. Proceedings of ACM Multimedia'09, pp.105–114.

[18] Jia, M.-L., Fan, X., Xie, X., Li, M.-J., and Ma, W.-Y. (2006) Photo-to-search: Using Camera Phones to Inquire of the Surrounding World. Mobile Data Management.

[19] Kleinberg, J. (1999) Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, Vol.46, No.5, pp.604–622.

[20] Lafferty, J., McCallum, A., and Pereira, F. (2001) Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th Conference on Machine Learning: pp.282–289.

[21] Mandl, T., Carvalho, P., Nunzio, G.M.D., Gey, F., Larson, R.R., Santos, D., Womser-Hacker, C. (2008) GeoCLEF 2008: The CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. Proceedings of CLEF 2008, pp.808–821.

[22] Mandl, T., Gey, F., Nunzio, G.M.D., Ferro, N., Larson, R.R., Sanderson, M., Santos, D., Womser-Hacker, C., Xie, X. (2007) GeoCLEF 2007: The CLEF 2007

Cross-Language Geographic Information Retrieval Track Overview. Proceedings of CLEF 2007, pp.745–772.

[23] Manning, C. D. and Schu tze, H (2000) Foundations of Statistical Natural Language Processing, chapter 7, MIT Press.

[24] Nakatoh, T., Yin, C., and Hirokawa, S. (2011) Characteristic Grammatical Context of Tourism Information,ICIC Express Letters, Vol.4, No.5.

[25] Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A., and Takezawa, T. (2009) Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, pp.205–208.

[26] Nister, D. and Stewenius, H. (2006) Scalable Recognition with a Vocabulary Tree. Proceedings of CVPR 2006.

[27] Ott, M., Choi, Y. Cardie, C., and Hancock, J.T. (2011) Finding Deceptive Opinion Spam by Any Stretch of the Imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp.309–319.

[28] Prantner, K. (2004) OnTour -The Ontology-, DERI Innsbruck.

[29] Quirk, C., Brockett, C., and Dolan, W. (2004) Monolingual Machine Translation for Paraphrase Generation. Proceedings of EMNLP 2004, pp.142–149.

[30] Ruiz-Martínez, J.M., Castellanos-Nieves, D., Valencia-Garcĩa, R., Fernãndez-Breis, J.T., Garcĩa-Sãnchez, F., Vivancos-Vicente, P.J., Castejõn-Garrido, J.S., Camõn, J.B., and Martĩnez-Bẽjar, R. (2009) Accessing Touristic Knowledge Bases through a Natural Language Interface, Proceedings of PKAW 2008, LNAI 5465, pp.147–160.

[31] Titov, I. and McDonald, R. (2008) A Joint Model of Text and Aspect Ratings for Sentiment Summarization. Proceedings of Annual Meeting of the Association for Computational Linguistics & Human Language Technology, pp.308–316.

[32] Tsai, R.T.-H. and Chou, C.-H. (2011) Extracting Dish Names from Chinese Blog Reviews Using Suffix Arrays and a Multi-Modal CRF Model, Proceedings of ACM SIGIR 2011.

[33] World Tourism Organization (2001) Thesaurus on Tourism and Leisure Activities of the World Tourism Organization.

[34] Wu, X., Li, J., and Neo, S.-Y. (2008) Personalized Multimedia Web Summarization for Tourist. Proceedings of World Wide Web Conference 2008.

[35] Zhao, S., Niu, C., Zhou, M., Liu, T., and Li, S. (2008) Combining Multiple Resources to Improve SMT-based Paraphrasing Model. Proceedings of ACL-HLT 2008, pp.1021–1029.

[36] Zhou, L., Lin, C.-Y., Munteanu, D.S., and Hovy, E. (2006) ParaEval: Using Paraphrases to Evaluate Summaries Automatically. Proceedings of HLT-NAACL 2006, pp.447–454.