Automatic Compilation of Travel Information from Automatically Identified Travel Blogs

Hidetsugu Nanba
Graduate School of Information Sciences, Hiroshima City University
nanba@hiroshima-cu.ac.jp

Haruka Taguma
School of Information Sciences, Hiroshima City University

Takahiro Ozaki
School of Information Sciences, Hiroshima City University

Daisuke Kobayashi
Graduate School of Information Sciences, Hiroshima City University
kobayashi@ls.info.hiroshima-cu.ac.jp

Aya Ishino
Graduate School of Information Sciences, Hiroshima City University
ishino@ls.info.hiroshima-cu.ac.jp

Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
takezawa@hiroshima-cu.ac.jp

Abstract

In this paper, we propose a method for compiling travel information automatically. For the compilation, we focus on travel blogs, which are defined as travel journals written by bloggers in diary form. We consider that travel blogs are a useful information source for obtaining travel information, because many bloggers' travel experiences are written in this form. Therefore, we identified travel blogs in a blog database and extracted travel information from them. We have confirmed the effectiveness of our method by experiment. For the identification of travel blogs, we obtained scores of 38.1% for Recall and 86.7% for Precision. In the extraction of travel information from travel blogs, we obtained 74.0% for Precision at the top 100 extracted local products, thereby confirming that travel blogs are a useful source of travel information.

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 them up to date manually. Therefore we have studied the automatic compilation of travel information.

For the compilation, we focused on travel blogs, which are defined as travel journals written by bloggers in diary form. Travel blogs are considered a useful information source for obtaining travel information, because many bloggers' travel experiences are written in this form. Therefore, we identified travel blogs in a blog database, and extracted travel information from them.

Travel information in travel blogs is also useful for recommending information that is matched to the each traveler. Recently, several methods that identify bloggers' attributes such as residential area (Yasuda *et al.*, 2006), gender, and age (Ikeda *et al.*, 2008, Schler *et al.*, 2006), have been proposed. By combining this research with travel information extracted from travel blogs, it is possible to recommend a local product that is popular among females, for example, or a travel spot, where young people often visit.

The remainder of this paper is organized as follows. Section 2 describes related work. Section 3 describes our method. To investigate the effectiveness of our method, we conducted some experiments, and Section 4 reports the experimental results. We present some conclusions in Section 5.

2 Related Work

Both 'www.travelblog.org' and 'travel.blogmura.com' are portal sites for travel blogs. At these sites, travel blogs are manually registered by bloggers themselves, and the blogs are classified by their destinations. However, there are many more travel blogs in the blogos-

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 205–208,
Suntec, Singapore, 4 August 2009.

phere. Aiming to construct an exhaustive database of travel blogs, we have studied the automatic identification of travel blogs.

GeoCLEF[1] is the cross-language geographic retrieval track run as part of the Cross Language Evaluation Forum (CLEF), and has been operating since 2005 (Gey *et al.*, 2005). The goal of this task was to retrieve news articles relevant to particular aspects of geographic information, such as 'wine regions around the rivers in Europe'. In our work, we focused on travel blogs instead of news articles, because bloggers' travel experiences tend to be written in travel blogs.

3 Automatic Compilation of Travel Information

The task of compiling travel information is divided into two steps: (1) identification of travel blogs and (2) extraction of travel information from them. We explain these steps in Sections 3.1 and 3.2.

3.1 Identification of Travel Blogs

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 writes his/her journey to Norway in multiple blog entries, it 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 that the second entry is a travel blog. Therefore, we focus not only on each entry but also on its surrounding entries for the identification of travel blogs.

We formulated the identification of travel blogs as a sequence-labeling problem, and solved it using machine learning. For the machine learning method, we examined the Conditional Random Fields (CRF) method, whose empirical success has been reported recently in the field of natural language processing. The CRF-based method identifies the class of each entry. Features and tags are given in the CRF method as follows: (1) the 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). We used the value of k=4, which was determined in a pilot study. Here, we used the following features for machine learning: whether an entry contains each 416 cue phrase, such as '旅行 (travel)', 'ツアー (tour)', and '出発 (departure)', and the number of location names in each entry[2].

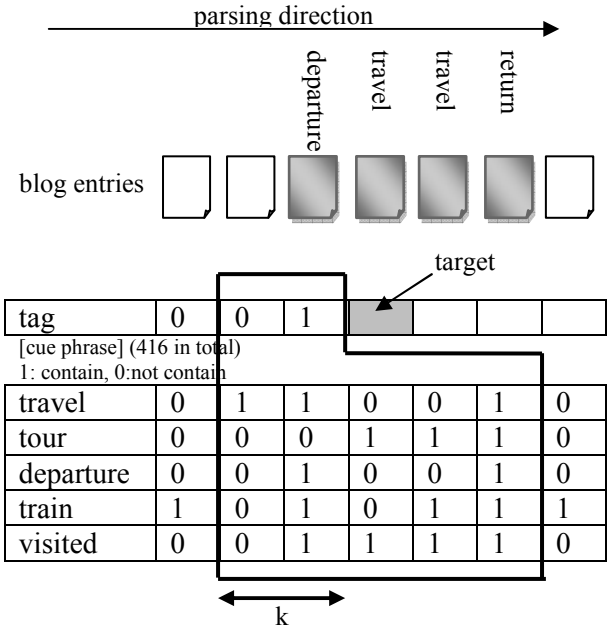

Figure 1: Features and tags given to the CRF

3.2 Extraction of Travel Information from Blogs

We extracted pairs comprising a location name and a local product from travel blogs, which were identified in the previous step. For the efficient extraction of travel information, we employed a bootstrapping method. Firstly, we prepared 482 location-name/and local-product pairs as seeds for the bootstrapping. These pairs were obtained automatically from a 'Web Japanese N-gram' database[3] provided by Google, Inc. The database comprises N-grams (N=1–7) extracted from 20 billion of Japanese sentences on the web. We applied a pattern '[地名]名物「[名物]」' ([slot of 'location name'] local product 「[slot of 'local product name']」) to the database, and extracted location names and local products from each corresponding slot, thereby obtaining the 482 pairs.

Secondly, we applied a machine learning-based information extraction technique to the travel blogs identified in the previous step, and obtained new pairs. In this step, we prepared

[1] http://ir.shef.ac.uk/geoclef/

[2] We used CaboCha software for the identification of locations.
http://chasen.org/~taku/software/cabocha/

[3] http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html

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 'location' and 'product' tags are assigned.
2. Prepare another 200 sentences that contain only a location name.[4] 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 annotates 'location' and 'product' tags to given sentences.

As a machine learning method, we used the CRF. In the same way as in the previous step, 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) the k tags occur before a target word, (2) k features occur before a target word, and (3) k features follow a target word. We used the value of k=2, which was determined in a pilot study. We use the following six features for machine learning.

- A word.
- Its part of speech[5].
- 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'.

4 Experiments

We conducted two experiments: (1) identification of travel blogs, and (2) extraction of travel information from blogs. We reported on them in Sections 4.1 and 4.2.

4.1 Identification of Travel Blogs

Data sets and experimental settings

We randomly selected 4,914 blog entries written by 317 authors from about 1,100,000 entries written in Japanese. Then we manually identified travel blogs in 4,914 entries. As a result, 420 entries were identified as travel blogs. Then we performed a four-fold cross-validation test. For the machine-learning package, we used CRF++[6] software. For evaluation measures, we used Recall and Precision scores.

Alternatives

In order to confirm the validity of our sequence labeling-based approach, we also examined another method, which identifies travel blogs using features in each blog entry only (without using features in its surrounding entries).

Results and discussions

Table 1 shows the experimental results. As shown in the table, our method improved the Precision value by 26.2%, while decreasing the Recall value by 13.0%. In our research, Precision is more important than Recall, because low Precision in this step causes low Precision in the next step.

	Recall	Precision
our method	38.1	86.7
baseline method	51.1	60.5

Table 1: Identification of travel blogs

Our method could not identify 266 of the travel blogs. We randomly selected 50 entries from these 266, and analysed the errors. Among the 50 errors, 25 cases (50%) were caused by the lack of cue phrases. For the machine learning, we used manually selected cue phrases. To increase the number of cue phrases, a statistical approach will be required. For example, applying n-grams to automatically identified travel blogs is one such approach. Among the 50 errors, 5 entries (10%) were too short (fewer than four sentences) to be identified by our method.

Our method mistakenly identified 26 entries as travel blogs. A typical error is that bloggers wrote non-travel entries among a series of travel blogs. In this case, the non-travel entries were identified as travel blogs.

4.2 Extraction of Travel Information from Blogs

Data sets and experimental settings

To confirm that travel blogs are a useful information source for the extraction of travel information, we extracted travel information using the following three information sources.

[4] In our pilot study, we did not use these negative cases in machine learning at first, and obtained low precision values, because our system attempted to extract local products from all sentences containing location names in travel blogs.

[5] In this step, we also identified location names automatically using the CaboCha software.

[6] http://www.chasen.org/~taku/software/CRF++/

- **Travel blogs (our method)**: 80,000 sentences in 17,268 travel blogs, which were automatically identified from 1,100,000 entries using the method described in Section 3.1.
- **Generic blogs**: 80,000 sentences from 1,100,000 blog entries.
- **Generic webs**: 80,000 sentences from 470M web sentences (Kawahara and Kurohashi, 2006).

We extracted travel information (location-name/local-product pairs) from each information source, and ranked them by their frequencies.

Evaluation

We used the Precision value for the top-ranked travel information defined by the following equation as the evaluation measure. We calculated Precision values from the top 5 to the top 100 at intervals of 5.

$$\text{Precision} = \frac{\text{The number of correctly extracted location-name / local-product pairs}}{\text{The number of extracted location-name / local-product pairs}}$$

Results and discussions

Figure 2 shows the experimental results. As shown in the figure, the generic blog method obtained higher Precision values than the generic web method, especially at higher ranks. Our method (travel blog) was much better than the generic blog method, which indicates that travel blogs are a useful information source for the extraction of travel information.

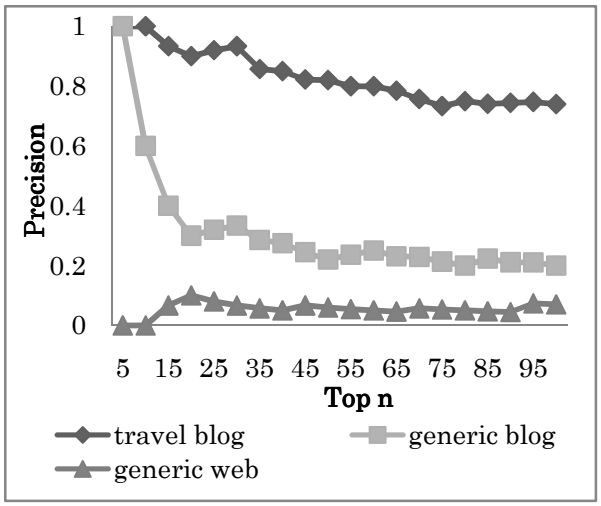

Figure 2: Precision values at top n for the extraction of travel information

Table 2 shows the number of local products, which were not contained in a list of products from the Google N-gram database. As shown in the table, 41 local products were newly extracted from travel blogs, while 15 and 7 were extracted from generic blogs and generic webs, respectively. These results also indicate the effectiveness of travel blogs as a source for travel information.

A typical error among the top 100 results for our method was that store names were mistakenly extracted. Here, most of these stores sell local products. To ameliorate this problem, extraction of pairs of local products and the stores that sell them is also required.

travel blog (our method)	41
generic blog	15
generic web	7

Table 2: The number of local products that each method newly extracted

5 Conclusion

In this paper, we proposed a method for identifying travel blogs from a blog database, and extracting travel information from them. In the identification of travel blogs, we obtained of 38.1% for Recall and 86.7% for Precision. In the extraction of travel information from travel blogs, we obtained 74.0% for Precision with the top 100 extracted local products.

References

Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. 2005. GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. *Lecture Notes in Computer Science*, LNCS4022, pp.908-919.

Daisuke Ikeda, Hiroya Takamura, and Manabu Okumura. 2008. Semi-Supervised Learning for Blog Classification. *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp.1156-1161.

Daisuke Kawahara and Sadao Kurohashi. 2006. A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp.176-183.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. *Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs*, pp.199-205.

Norihito Yasuda, Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki. 2006. Identifying bloggers' residential areas. *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp.231-236.