Automatic Identification of Know-How Blog Entries from a Travel Blog Database

Hidetsugu Nanba Saki Douke Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
Hiroshima, Japan
{nanba, douke, takezawa}@ls.info.hiroshima-cu.ac.jp

***Abstract*—Tourists in foreign countries must keep in mind differences in culture and customs of the country they are visiting to avoid any trouble that may arise from these differences and to enjoy and appreciate the culture. We propose a method for identifying blogs that offer travel know-how from a travel blog database. To investigate the effectiveness of our method, we conducted some experiments. From the experimental results, we obtained precision of 0.619, recall of 0.928, and F-measure of 0.743.**

Keywords-component; blog; travel know-how; SVM

I. INTRODUCTION

Tourists visiting foreign countries must keep in mind differences in culture and customs of the new country to help them avoid any trouble that may arise from these differences. In this paper, we propose a method to identify blogs containing travel know-how from a travel blog database.

The travel know-how blog entries can be divided into the following two types:

1. Event-type entries: These entries mention unique travel-related events, such as a traditional Japanese tea ceremony. These blog entries give information about the backgrounds or history of the events, manners including dress code for attending the events, how to attend, and time required for the events.
2. Procedure-type entries: Examples are "how to climb Mt. Fuji" and "how to put on a traditional Japanese kimono."

Acquiring travel know-how through automatically identified know-how blog entries is useful not only for avoiding trouble caused by a lack of information but also for enjoying and appreciating the culture of a country to be visited.

The remainder of this paper is organized as follows. Section II describes related work. Section III explains our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section IV reports on these and the results. Section V concludes.

II. RELATED WORK

Kozawa et al. [1] proposed a method for acquiring know-how information by focusing on each object and how it is used. For example, when object names, such as "hair dryer" or "thermometer," are given to their system, it outputs the corresponding know-how information, such as "to heat something" and "to monitor a temperature", respectively, which are automatically extracted from the Web. In contrast to their approach, we focus on travel, and identify travel know-how blog entries that mention about unique events and procedures as described in Section I.

Inui et al. [2] proposed a method for collecting instances of personal experiences as well as opinions from blogs. An example of such a personal experience is the following.

> *On my way home, I (in a wheelchair) could not find my way out of Totsuka Station because all the elevators in the station building stop running at 11 pm.*

Such personal experiences are considered useful for tourists. However, we focus on not only personal experiences but also knowledge including historical or well-known facts related to some events.

III. IDENTIFICATION OF KNOW-HOW BLOGS FROM A TRAVEL BLOG DATABASE

We use several cue phrases for identifying know-how blog entries. We now describe how to collect these cue phrases.

A. Manual Method

To collect cue phrases, we used a travel blog database "TravelBlog"[1], which provides more than 600,000 travel blog entries written in English. Among these entries, we randomly selected 20,000 entries containing the phrase "how to." We then investigated features of know-how blog entries, and found that five types of cue phrases in them. These are illustrated with some examples of cue phrases as follows.

- Methodology: how to make, how to get, how to eat
- Procedure: Step 1, step one, 1)
- Schedule: annual, national holiday
- Venue: festival in
- Action: take place, attend

B. Semiautomatic Method

To collect more cue phrases, we focused on the expression "how to." Some expressions, such as "how to eat" and "how to attend," are useful for identifying know-how blog entries, while some expressions, such as "how to be" and "how to explain" may not be useful for identification. We therefore collected candidates of cue phrases by applying N-gram statistics to the 20,000 blog entries above, and manually selected cue phrases from the candidates. Finally, we obtained 54 cue phrases in total, a few of which are shown in Table I.

[1]http://www.travelblog.org

TABLE I. EXAMPLE OF CUE PHRASES FOR TRAVEL KNOW-HOW BLOG IDENTIFICATION

how to get to the	how to play
how to cook	how to eat
how to make	how to buy

IV. EXPERIMENTS

To investigate the effectiveness of our methods, we conducted several experiments.

A. Datasets and Experimental Setting

To generate the test data for identifying know-how and event blog entries, we used the 426 blog entries containing the phrase "how to" in "TravelBlog." Then, we manually identified 222 know-how blog entries. For the identification, we employed the following criterion.

Even if a blog entry contains know-how information, we do not identify it as a know-how blog entry if more than half of it is irrelevant description.

B. Machine Learning and Evaluation Measures

We performed a twofold cross-validation test. We used TinySVM[2] as the machine-learning package and used a polynomial kernel of degree two. As evaluation measures, we used precision, recall, and F-measure.

C. Alternatives

We conducted tests using the following three methods and a baseline method.

Our methods

- CUE: Use cue phrases, such as those in Section III-A as features for machine learning.
- N-GRAM: Use cue phrases such as those in Table I in Section III-B as features for machine learning.
- CUE+N-GRAM: Use the cue phrases in Sections III-A and III-B as features for machine learning.

Baseline method

- BASE: Identify all blog entries as know-how blog entries.

D. Results and Discussion

Table II shows the experimental results. The F-measure score by our method CUE was 0.058 higher than that of the baseline method, confirming the effectiveness of our method for identifying know-how blog entries.

TABLE II. IDENTIFICATION OF TRAVEL KNOW-HOW BLOG ENTRIES

	Precision	**Recall**	**F-measure**
CUE	0.619	0.928	0.743
N-GRAM	0.558	0.946	0.702
CUE+N-GRAM	0.610	0.881	0.721
BASE	0.521	1.000	0.685

[2] http://chasen.org/~taku/software/TinySVM/

Generally, our methods could obtain high recall values, while precision values are low. We therefore discuss the low precision values. Figure 1 is a typical entry, which our method CUE mistakenly identified as a know-how blog entry. In this entry, two cue phrases appear (underlined), and as a result, CUE identified it as a know-how blog. This entry comprises two subtopics: (1) A ninja show, and (2) a coffee shop that the blogger visited after the ninja show. The blog thus disobeys the criteria in Section IV-A. This strict criterion is the main reason for the low precision value. In our future work, we will investigate the automatic extraction of travel know-how passages from each entry that is identified by our method.

Title: Ninja Show
On my final day in Japan, I hung out in Iga and finally got to watch the Ninja show. It was amazing watching ninja's fight each other, demonstrate how to use weapons, and even have kids come up to hold katanas.
(snip)
The younger ninja demonstrated how to throw one, two, and three shurikan at once.
(snip)
Aside from the ninja show we discovered a new coffee shop where we enjoyed a pastry for "lunch".
(snip)

Figure 1. Example of mistakenly identified blog entry.

V. CONCLUSION

In this paper, we proposed a method for identifying know-how blog entries in a travel blog database. From the experimental results, our method CUE obtained precision of 0.619, recall of 0.928 and F-measure of 0.743, confirming the effectiveness of our method.

REFERENCES

[1] S. Kozawa, K. Uchimoto, and S. Matsubara, "Acquisition of Know-How Information from Web," Lecture Notes in Computer Science, Vol. 7097, pp.446-457, 2011.

[2] K. Inui, S. Abe, H. Morita, M. Eguchi, A. Sumida, C. Sao, K. Hara, K. Murakami, and S. Matsuyoshi, "Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents," Proc. the 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp.314-321, 2008.