Automatic Extraction of Event Information from Newspaper Articles and Web Pages

Hidetsugu Nanba, Ryuta Saito, Aya Ishino, and Toshiyuki Takezawa

Hiroshima City University, Graduate School of Information Sciences,
3-4-1, Ozukahigashi, Asaminamiku, Hiroshima 731-3194 Japan

Abstract. In this paper, we propose a method for extracting travel-related event information, such as an event name or a schedule from automatically identified newspaper articles, in which particular events are mentioned. We analyze news corpora using our method, extracting venue names from them. We then find web pages that refer to event schedules for these venues. To confirm the effectiveness of our method, we conducted several experiments. From the experimental results, we obtained a precision of 91.5% and a recall of 75.9% for the automatic extraction of event information from news articles, and a precision of 90.8% and a recall of 52.8% for the automatic identification of event-related web pages.

Keywords: newspaper article, web page, event, travel.

1 Introduction

Information about events that are scheduled to take place at the travel destination is crucial when planning a trip. Travel guidebooks and portal sites provided by tour companies and government tourist offices are useful for checking on well-known events, such as festivals. However, it is costly and time consuming to compile information about the full range of events for all tourist spots manually and to keep this data up-to-date. We have therefore investigated the automatic compilation of event information.

We use newspaper articles and web pages as information sources for the extraction of event information. In general, we can obtain information about popular or traditional events through newspaper articles. However, only a small percentage of the full range of travel-related events appears in newspaper articles. Therefore, we propose a method to collect travel-related event information from both newspaper articles and web pages. Generally, public venues, such as museums or exhibition halls, have their own web sites, which contain web pages that give schedules for events at these venues (we call them "event web pages"). We therefore extract venue names, where public events often occur from news corpora, and then identify event web pages using this list. Using this procedure, we can expect to obtain much travel-related event information rapidly.

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

S.R. Urs, J.-C. Na, and G. Buchanan (Eds.): ICADL 2013, LNCS 8279, pp. 171–175, 2013.

method, we conducted experiments, with Section 4 reporting the results. We present some conclusions in Section 5.

2 Related Work

In previous work on extracting event information from texts, Schneider [2] and Issertial et al. [1] proposed a method for extracting academic event information, such as schedules, conference locations, and conference names using "Call for Papers" documents as information sources. In our work, we use news corpora and web pages as our information sources.

3 Construction of the Event-Retrieval System

We obtain the event information in two steps: (1) extraction of event information from newspaper articles and (2) identification of event web pages from the web. We describe these two steps in Sections 3.1 and 3.2, respectively.

3.1 Extraction of Event Information from Newspaper Articles

We use information extraction based on machine learning to extract event information from event news articles. First, we define the tags used:

- **EVENT** tag includes an event name;
- **DATE** tag includes a schedule for an event;
- **ADDRESS** tag includes an address of a venue;
- **LOCATION** tag includes a venue name.

We formulate the identification of the class of each word in a given sentence by using machine learning. We opted for the CRF machine-learning method, where the features and tags used are as follows: (1) k features occur before a target word and (2) k features follow a target word. We use the value $k = 3$, which was determined via a pilot study. We use the following features for machine learning. (We use MeCab as a Japanese morphological analysis tool to identify the part of speech):

- a word;
- its part of speech;
- whether the word is a quotation mark;
- whether the word is frequently used in the event names, such as "live", "festival", or "fair";
- whether the word is frequently used in the names of venues, such as "gallery", "stage", "building", or "hot spring";
- whether the word is a numerical descriptor;
- whether the word is frequently used in the address, such as "prefecture" or "city";
- whether the word is frequently used in schedules, such as "date", "schedule", or "tomorrow".

3.2 Identification of Event Web Pages

For the identification of event web pages, we first search web pages using pairs comprising a venue name[1] and the word "ibento" (event) as the query. We next identify event web pages using cue words and unnecessary words as features for machine learning. Examples of the use cue phrases and unnecessary words are:

- whether the web page includes cue words, such as "schedule", "date", "participation fee", "event", "information", or "fair";
- whether the web page includes unnecessary words, such as "youtube", "blog", "twitter", "livedoor", "facebook", "wikipedia", or "mapple";
- whether the web page's URL includes cue words, such as "event" or "schedule";
- whether the URL includes unnecessary words, such as "youtube" or "facebook";
- whether the web page includes a table.

4 Experiments

We conducted two experiments to test (1) the extraction of event information from news articles, and (2) the identification of event web pages.

4.1 Extraction of Event Information from News Articles

Data Sets and Experimental Settings

We used 416 event newspaper articles with four kinds of tags, namely EVENT, DATA, ADDRESS, and LOCATION, annotated manually.

Machine Learning and Evaluation Measures

We employed CRF as our machine learning technique, and performed a four-fold cross-validation test. As a baseline method, we used words and their parts of speech as features for machine learning. As evaluation measures, we again used recall, precision, and F-measure.

Experimental Results

We show the experimental results in Table 1. To compare with the performance of our method, we calculated scores of recall, precision, and F-measure, when extracting the whole articles as event news articles.

As shown in the table, our method improved the recall score by more than 0.1 without impairing the precision score. However, our method was unable to extract event information in 24.1% of cases, caused mainly by the lack of linguistic clues. Table 2 shows examples of errors when our method failed. In the EVENT error, "HARA Asao ten" (Hara Asao exhibit) should have been extracted, but our method failed, because no cue words, such as quotation marks or particles (topic marker), except for "ten" (exhibit) appear in this context. In the ADDRESS error, there are no linguistic clues around the location name "Umeda", which should have been extracted as an address.

[1] We obtained 30,000 venue names from news corpora automatically using the method of Section 3.1.

Table 1. Evaluation results for the extraction of event information from news articles

	Baseline			Our method		
	Precision	Recall	F-measure	Precision	Recall	F-measure
EVENT	0.899	0.510	0.651	0.912	0.673	0.774
DATE	0.965	0.799	0.874	0.968	0.855	0.908
ADDRESS	0.910	0.769	0.833	0.908	0.816	0.860
LOCATION	0.896	0.547	0.679	0.873	0.693	0.773
Average	0.918	0.656	0.765	0.915	0.759	0.830

Table 2. Examples of errors in the extraction of event information from event articles

EVENT error	kajin, HARA Asao ten deha... (at the poet HARA Asao exhibit...)
ADDRESS error	2/1 made, Daimaru * Umeda no 11 F (on the 11th floor of Daimaru (department store) * Umeda, until Feb. 1.)

4.2 Identification of Event Web Pages

Data Sets and Experimental Settings

We used 1,022 web pages containing the word "ibento" (event) for our experiment. Of these, 264 web pages were identified manually as event web pages.

Machine Learning and Evaluation Measures

We performed a four-fold cross-validation test, again using TinySVM software as the machine-learning package. We again used recall, precision, and F-measure as evaluation measures.

Experimental Results

We show the experimental results in Table 3. As baseline results to compare with those of our method, we calculated scores for recall, precision, and F-measure, when extracting all pages as event web pages.

Our method achieved the high precision score of 0.824, while the recall score remained at 0.522. In many of the event web pages that our method could not identify, event schedules were written in a tabular form. Although our method does check for web page containing tables, as one of the features for machine learning, these pages did not contain additional linguistic cues, which our method requires for proper identification.

Table 3. Evaluation results for identification of event web pages

Methods	Precision	Recall	F-measure
Baseline	0.258	1.000	0.410
Our method	0.824	0.522	0.639

5 Conclusions

We have proposed a method for extracting event information from both newspaper articles and event web pages. In the extraction of event information from news articles, we obtained a precision of 91.5% and a recall of 75.9%. For the identification of event web pages, we obtained a precision of 90.8% and a recall of 52.8%. Our method outperformed a baseline method, thereby confirming the effectiveness of our method.

References

1. Issertial, L., Tsuji, H.: Information Extraction and Ontology Model for a 'Call for Paper' Manager. In: Proceedings of iiWAS 2011, pp. 539–542 (2011)
2. Schneider, K.-M.: An Evaluation of Layout Features for Information Extraction from Calls for Papers. In: LWA 2005, pp. 111–115 (2005)