Extracting Transportation Information and Traffic Problems from Tweets during a Disaster

Where do you evacuate to?

Aya Ishino, Shuhei Odawara, Hidetsugu Nanba, Toshiyuki Takezawa
Graduate School of Information Sciences, Hiroshima City University
Hiroshima, Japan
{ishino, odawara, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

***Abstract*—In a disaster, one of the most important issues for victims is how to find evacuation routes to safety from hazardous areas. To offer such routes, we propose methods automatically extracting transportation information and traffic problems from tweets written in Japanese and posted during a disaster. To investigate the effectiveness of our methods, we conducted some experiments using tweets posted during the Great Eastern Japan Earthquake in March 2011. From the experimental results, we obtained precision of 78.2% and recall of 53.4% in automatic extraction of transportation information. For extracting traffic problems, we identified tweets containing relevant information (we call them traffic problem tweets), and extracted traffic problem from them. In identifying traffic problem tweets, we obtained precision of 77.7% and recall of 70.7%. In extracting traffic problems, we obtained precision of 87.0% and recall of 57.1%. Thus, we have constructed a system for providing transportation information and traffic problems in a disaster.**

Keywords-disaster; evacuation routes; information extraction.

I. INTRODUCTION

Disasters occur frequently throughout the world. For instance, there were the large earthquakes in Haiti in January 2010 and in Sumatra, Indonesia in December 2004. In March 2011, a massive earthquake of magnitude 9.0 struck off the coast of eastern Japan. This earthquake is called the Great Eastern Japan Earthquake. It caused tsunamis and an accident at a nuclear power plant, and forced large numbers of people to evacuate from their homes and towns. In such disasters, one of the most important issues for victims is how to find evacuation routes to safety from hazardous areas. To offer such evacuation routes, there is a need to collect transportation information and traffic problems from other victims. Because they are so widely used, we focused on extracting information from tweets.

After the Great Eastern Japan Earthquake, 18 million tweets were posted on Twitter in a day, which is 1.8 times as much as normal. Some tweets contained useful information about transportation and traffic problems. In this paper, we propose methods for extracting transportation information and traffic problems automatically from tweets posted during disasters. In addition, we construct a system for presenting the extracted information. We believe that the system can offer safe evacuation routes for disaster victims and transportation routes for relief materiel.

The remainder of this paper is organized as follows. Section II describes the system behavior using snapshots. Section III describes related work. Section IV explains our methods. To investigate the effectiveness of our methods, we conducted some experiments, and Section V reports on these and the results. We present some conclusions in Section VI.

II. SYSTEM BEHAVIOR

In this section, we describe our prototype system, which (1) provides transportation information, and (2) identifies traffic problems. Fig. 1 shows transportation information from disaster victims. Arrows with icons indicate transportation information. The arrow extends from a departure place (shown as ① in Fig. 1) to a destination (②). Each icon depicts a transportation method. If the user clicks the icon (③), the system shows details of transportation information (④). The user can discover that a disaster victim evacuated from Ishinomaki City to Ichinoseki City by car.

Fig. 2 shows traffic problems. An arrow with an icon indicates a traffic problem. A traffic problem is indicated by an arrow with an icon. The arrow shows that a traffic problem has occurred between one end of the arrow (shown as ① in Fig. 2) and the other end the arrow (②). If the user clicks the icon (③), the system shows details of the traffic problem (④). The user can discover that a traffic problem has occurred between Sendai City and Yamagata City on Route 48. In this paper, we describe the methods used by our system for extracting transportation information and traffic problems.

III. RELATED WORK

In this section, we describe some related studies on information mining in disasters and extracting transportation information.

A. Information Mining in Disasters

In time of disaster, vast amounts of data are generated via computer-mediated communication; however, it is difficult to extract useful information from them for users. There have been some studies of information mining in several disasters.

Verma *et al*. [1] collected tweets from four different disasters, and automatically detected tweets that could contribute to situational awareness automatically. They obtained over 80% accuracy.

Sakaki *et al.* [2] considered each Twitter user as a sensor, and detected disaster events based on sensory observation.

Copyright (c) IARIA, 2012. **ISBN: 978-1-61208-227-1**

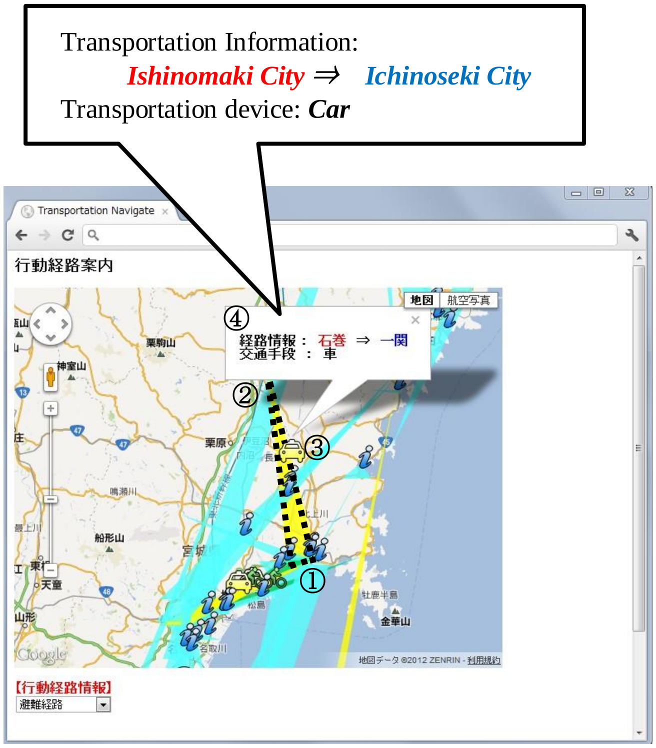

Figure 1. The system that provides transportation information.

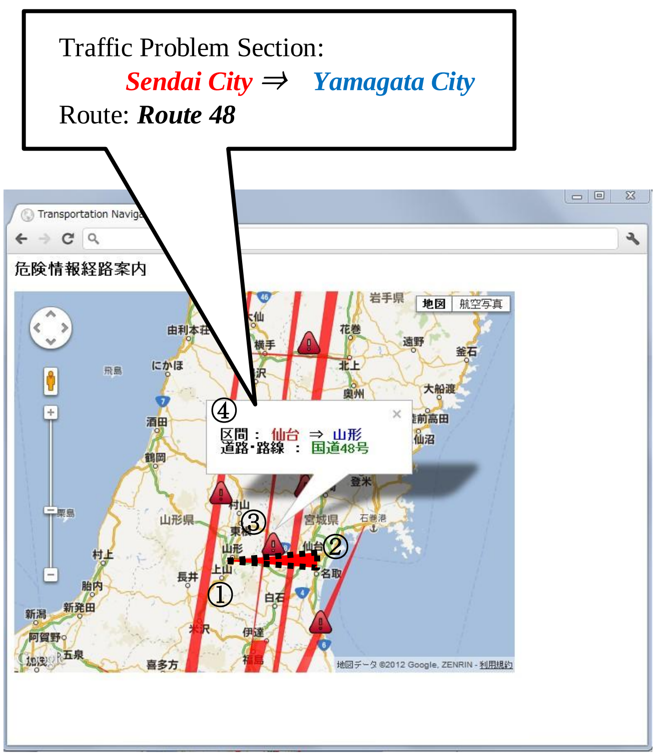

Figure 2. The system that provides traffic problem.

They targeted disaster events such as earthquakes and typhoons. As an application, they constructed an earthquake reporting system.

There are some studies about evacuation in disasters. Iwanaga *et al.* [3] build an earthquake evacuation ontology from twitter and provided the most suitable evacuation center. Troung *et al.* [4] presented a novel framework that manages and provides various types of context information required for adapting processes in emergency management systems.

Soon after the Great Eastern Japan Earthquake, many Natural Language Processing (NLP) researchers, engineers, and students from all over Japan created a working group, called "ANPI_NLP." "ANPI" means "safety" in Japanese. ANPI_NLP tried to collect tweets with hash tags, such as "#anpi (safety information)" or "#hinan (evacuation)", and extracted information about the safety of people. [5] In this paper, we use the tweet corpus provided by ANPI_NLP, and extract transportation information and traffic problems from it.

B. Extracting Transportation Information

There have been a number of studies of extracting transportation information. Davidov [6] presented an algorithm framework that enabled automated acquisition of map-link information from the Web based on surface patterns such as "from X to Y." Given a set of locations as initial seeds, they retrieved an extended set of locations from the Web and produced a map-link network that connected these locations using transport-type edges. In this paper, we propose a method for extraction of transportation information via machine-learning techniques.

Ishino *et al.* [7] extracted traveler's transportation information automatically from travel blog entries written in Japanese using machine-learning techniques. They used cues related to travel, such as "観光" (sightseeing tour) or "旅行" (travel) for machine learning. In this paper, we aim to extract transportation information from disaster victims. Therefore, we collect cues related to disasters for machine learning.

IV. EXTRACTING TRANSPORTATION INFORMATION AND TRAFFIC PROBLEMS

In this paper, we propose methods for extracting transportation information and traffic problems from tweets written in Japanese and posted during the Great Eastern Japan Earthquake. We explain our methods for extracting transportation information in Section IV-A, and for traffic problems in Sections IV-B and IV-C.

A. Extracting Transportation Information

In this section, we describe our method for extracting transportation information from tweets. We use machine learning to extract information, such as "a departure place", "a destination", or "a transportation method", from tweets. First, we define the tags used in our examination. Fig. 3 is a tagged example.

- FROM tag includes a departure place.
- TO tag includes a destination.
- METHOD tag includes a transportation method.

[Original]
(Tweet 1)
私の祖母は<FROM>山田岡</FROM>在住です。津波被害はなくガラスが数枚割れたと聞きました。避難勧告を受けて、楢葉の<METHOD>バス</METHOD>で<TO>いわき市の草野中学校</TO>に避難しています。
(Tweet 2)
義弟の安否確認が取れました。<FROM>石巻</FROM>から<METHOD>徒歩</METHOD>で<TO>仙台市内</TO>の家まで帰って来たそうです。

[Translation]
(Tweet 1)
My grandmother lives in <FROM>Yamadaoka</FROM>. The tsunami caused little damage there. She was urged to evacuate, and went to <TO>Kusano junior high school in Iwaki City</TO> by <METHOD>bus</METHOD>.
(Tweet 2)
I found out my brother-in-law is safe. He came back home in <TO>Sendai City</TO> from <FROM>Ishinomaki City</FROM> on <METHOD>foot</METHOD>.

Figure 3. Examples of tagged tweets.

We formulate the identification of the class of each word in a given sentence and solve it using machine learning. For the machine-learning method, we opted for the Conditional Random Fields (CRF) method [8]; its empirical success has been reported recently in natural language processing. 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. We used the value k = 6, which was determined via a pilot study. We used the following 13 features for machine learning. A sequence of nouns (a noun phrase) was treated as a noun. We used MeCab [9] 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 a cue phrase, as shown in Table I.

B. Identifying Traffic Problem Tweets

We proposed a method for extracting traffic problems from tweets posted during the Great Eastern Japan Earthquake. In a pilot study, we investigated the number of tweets containing traffic problem information (we call them traffic problem tweets) and found some examples. Therefore, this task is divided into two steps: (1) identifying traffic problem tweets from a tweet corpus; and (2) extracting traffic problems from the traffic problem tweets. We explain Step 1 in this section and Step 2 in Section IV-C.

In this section, we explain our method for identification of traffic problem tweets automatically. Fig. 4 shows examples of traffic problem tweets. They contain words related to traffic problems, such as "通行止め" (closed to traffic) or "停止" (shut down), and the name of the relevant road. We employed Support Vector Machine (SVM) [10] as a machine-learning technique to identify traffic problem tweets. We use the following features for machine learning.

TABLE I. CUE PHASES FOR EXTRACTION OF TRANSPORTATION INFORMATION

Tag	Cue phase	The number of cues
FROM	Whether the word is a cue that often appears immediately after the FROM tag, such as "から" (from) or "を出発" (left).	5
FROM TO	Whether the word is frequently used in the name of a shelter, such as "学校" (school) or "公民館" (community center).	23
	Whether the word is a cue that the FROM tag and the TO tag do not contain, such as "方向" (directions) or "沿い" (along).	7
	Whether the word is the name of a station, provided by ANPI_NLP.	8619
	Whether the word is the spot name in eastern Japan, provided by ANPI_NLP.	1755
	Whether the word is the name of a school in eastern Japan, provided by ANPI_NLP.	806
	Whether the word is the name of a train line, provided by ANPI_NLP.	569
TO	Whether the word is a cue that often appears immediately after the TO tag, such as "まで" (to) or "へ避難" (evacuate).	30
METHOD	Whether the word is a cue that often appears immediately after the METHOD tag, such as "で行く" (by).	19
	Whether the word is the name of a transportation device, such as "飛行機" (airplane) or "自動車" (car).	37

- The word relates to traffic problem, such as “通行止め” (closed to traffic) or “停止” (shut down) (19).
- The word relates to a road, such as “自動車道” (Expressway) or “インターチェンジ” (Interchange) (13).
- The word relates to transportation devices, such as “新幹線” (Shinkansen bullet train) or “地下鉄” (subway) (9).

C. *Extracting Traffic Problems*

In this section, we explain our method for extracting traffic problems from traffic problem tweets identified in Section IV-B. We use machine learning to extract information such as “a road” or “a train line”, or “traffic problem section”, from tweets. First, we define the tags used in our examination. Fig. 5 is a tagged example.

- LINE tag includes a road or a train line.
- LOC tag includes a traffic problem section.

We use CRF for the machine learning. 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. We used the value k = 6, which was determined via a pilot study. We use the following 14 features for machine learning. A sequence of nouns (a noun phrase) was treated as a noun. We used MeCab as a Japanese morphological analysis tool.

- A word.
- Its part of speech.
- Whether the word is a quotation mark.
- Whether the word is a mark, such as “~”, or “→”.
- Whether the word is a cue phrase, as shown in Table II.

[Original]
(Tweet 1)
地震で中央自動車道も上野原—勝沼インターチェンジ間などが通行止め。
(Tweet 2)
新幹線 浜松～品川停止中。

[Translation]
(Tweet 1)
After a large earthquake, the Chuo Expressway is closed to traffic between Uenohara city and the Katsunuma Interchange.
(Tweet 2)
Shinkansen (Bullet Train) are shout down. Hamamatsu – Shinagawa

Figure 4. Example of traffic problem tweets.

[Original]
(Tweet 1)
地震で<LINE>中央自動車道</LINE>も<LOC>上野原</LOC>—<LOC>勝沼インターチェンジ</LOC>間などが通行止め。
(Tweet 2)
<LINE>新幹線</LINE> <LOC>浜松</LOC>～<LOC>品川</LOC>停止中。

[Translation]
(Tweet 1)
After a large earthquake, <LINE>the Chuo Expressway</LINE> is closed to traffic between <LOC>Uenohara city</LOC> and <LOC>the Katsunuma Interchange</LOC>.
(Tweet 2)
<LINE>Shinkansen (Bullet Train)</LINE> are shout down. <LOC>Hamamatsu</LOC> – <LOC>Shinagawa</LOC>

Figure 5. Example of tagged traffic problem tweets.

TABLE II. CUE PHASE FRO EXTRACTION OF TRAFFIC PROBLEMS

Tag	Cue phase	The number of cues
LINE	Whether the word is frequently used in the name of a road and a train line, such as “道路” (road) or “号線” (line).	23
	Whether the word is the name of a train line, provided by ANPI_NLP.	569
	Whether the word is the name of a bypass, collected from Wikipedia.	1301
	Whether the word is the name of an express way.	60
	Whether the word is the name of a toll road.	181
LINE LOC	Whether the word frequently used in traffic problems, such as “通行止め” (closed to traffic) or “停止” (shout down) .	51
	Whether the word frequently used in traffic problems, such as “通行可能” (available for traffic) or “復旧” (restoration) .	19
LOC	Whether the word is the name of a station, provided by ANPI_NLP.	8619
	Whether the word is the spot name in eastern Japan, provided by ANPI_NLP.	1755
	Whether the word is a cue that often appears in the LOC tag, such as “駅” (station) or “インターチェンジ” (interchange).	10

Copyright (c) IARIA, 2012.
ISBN: 978-1-61208-227-1

V. Experiments

A. Ectracting Transportation Information

Data Sets and Experimental Settings

We randomly selected 1303 tweets written in Japanese from the tweet corpus provided by ANPI_NLP, and tagged them manually, as described in Section IV-A. The numbers of manually assigned tags are shown in Table III. We used CRF++ [11] software as the machine-learning package. As a base line method, we used only a word as a feature for machine learning. We used precision and recall as evaluation measures, calculated as follows.

$$\text{Precision} = \frac{\text{The number of correctly extracted tags}}{\text{The number of tags that the system extracted}} \quad (1)$$

$$\text{Recall} = \frac{\text{The number of correctly extracted tags}}{\text{The number of tags that should be extracted}} \quad (2)$$

TABLE III. Numbers of Manually Assigned Tags in The Extracted Transportation Information

Tag	Training	Test
FROM	237	71
TO	425	120
METHOD	61	17
Total	723	208

Results and Discussion

The evaluation results are shown in Table IV. Our method obtained higher precision and recall than the baseline method. We discuss the experimental results as follows.

TABLE IV. Evaluation Results for Extracting Transportation Information

Tag	Our method		Baseline method	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
FROM	78.4	40.9	72.4	29.6
TO	76.3	59.2	73.2	43.3
METHOD	91.7	64.7	80.0	23.5
Total	78.2	53.4	73.3	37.0

[Original]
(Correct) 助川小学校に避難された方がいらっしゃいましたら、現在どのような状況か情報頂きたいです！
(Analysis result) <TO>助川小学校</TO>に避難された方がいらっしゃいましたら、現在どのような状況か情報頂きたいです！

[Translation]
(Correct) If victims evacuate to Sukegawa Elementary School, please let me know what is going on!
(Analysis result) If victims evacuate to <TO>Sukegawa Elementary School</TO>, please let me know what is going on!

Figure 6. Example of a failure in extracting transportation information.

First, we discuss a typical error causing low precision. Fig. 6 shows an example of a failure in extracting transportation information. The TO tag was mistakenly assigned to "助川小学校" (Sukegawa Elementary School), which might not be an actual evacuation site. This was because the "TO" cue "避難" (evacuate) appears immediately before it. To improve the performance of extracting transportation information, we should consider language structure.

Next, we discuss a typical error causing low recall. A typical error is the lack of cues. In particular, we could not collect the names of some facilities or places cyclopedically. When preparing for a disaster, we must collect the names of facilities and places all over the world.

B. Identifying Traffic Problem Tweets

Data Sets and Experimental Settings

For our examination, we identified traffic problem tweets among 1750 tweets written in Japanese provided by ANPI_NLP. The number of manually identified traffic problem tweets is shown in Table V. We performed a four-fold cross validation test. We used a standard SVM package, TinySVM (http://chasen.org/~taku/software/TinySVM/). We used precision and recall as evaluation measures.

TABLE V. Number of Manually Identified Traffic Problem Tweets

Traffic Problem Tweets	Others	Total
166	1584	1750

Results and Discussion

Table VI shows the experimental results. Our method obtained a higher recall than baseline method. We now discuss the low recall of our method. A typical reason for low recall is the lack of cues. For machine learning, we used manually selected cues, as described in Section IV-B. To improve the coverage of cues, a statistical approach, such as applying n-gram statistics to a larger tweet corpus, will be required.

TABLE VI. Evaluation Results for Identifying Traffic Problem Tweets

	Precision (%)	Recall (%)
Our method	77.7	70.7
Baseline method	80.4	61.9

C. Extracting Traffic Problems

Data Sets and Experimental Settings

We manually assigned tags to traffic problem tweets, as described in Section IV-C, and used them for our examination. Table VII shows the numbers of manually assigned tags. We used CRF++ software as the machine-learning package. As a baseline method, we used only a word as a feature for machine learning. We used recall and precision as evaluation measures, calculated as equations (1) and (2).

TABLE VII. NUMBERS OF MANUALLY ASSIGNED TAGS FOR EXTRACTING TRAFFIC PROBLEMS

Tag	Training	Test
LINE	126	39
LOC	176	67
Total	166	243

Results and Discussion

The evaluation results are shown in Table VIII. Our method obtained higher recall than the baseline method.

We now discuss the low recall of our method. Errors in our method have been found in tweets that contain both problem information and safety information about traffic states. Fig. 7 shows an example of failure in the extracting traffic problems. In the example, the "LINE" tag should be assigned to "県内高速道路" (expressway in the prefecture), but our method did not assign any tags to this word. This tweet contains the cue "通行止め" (closed to traffic), and the cue "通行可能" (can pass). In this case, we should consider language modification relationships of cues.

TABLE VIII. EVALUATION RESULTS FOR EXTRACTING TRAFFIC PROBLEMS

Tag	Our method		Baseline method	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
LINE	89.7	68.4	80.0	31.6
LOC	85.0	50.8	92.0	34.3
Average	87.0	57.1	87.5	33.3

[Original]
(Correct) 能代南ＩＣ－二ツ井ＩＣ間は通行可能。そのほかの<LINE>県内高速道路</LINE>通行止め。
(Analysis result) 能代南ＩＣ－二ツ井ＩＣ間は通行可能。そのほかの<line>県内高速道路</line>通行止め。

[Translation]
(Correct) You can pass between the Noshiro Minami Interchange and the Futatsui Interchange. Another <LINE>expressway in the prefecture</LINE> is closed to traffic.
(Analysis result) You can pass between the Noshiro Minami Interchange and the Futatsui Interchange. Another expressway in the prefecture is closed to traffic.

Figure 7. Example of a failure in extracting traffic problems.

VI. CONCLUSION

To offer evacuation routes to safety for disaster victims, we have proposed methods for extracting transportation information and traffic problems from tweets posted during disasters. To investigate the effectiveness of our methods, we conducted some experiments using tweets posted during the Great Eastern Japan Earthquake. From the experimental results, we obtained precision of 78.2% and recall of 53.4% in automatic extraction of transportation information. For extracting traffic problems, we identified tweets containing information about traffic problems (we called them traffic problem tweets), and extracted traffic problems from them. In identifying traffic problem tweets, we obtained precision of 77.7% and recall of 70.7%. In extracting traffic problems, we obtained precision of 87.0% and recall of 57.1%. Thus, we have constructed a system for providing transportation information and identifying traffic problems in disasters. We consider that the system can offer evacuation routes for disaster victims and transportation routes for relief materiel.

In this paper, we used tweets that posted during the Great Eastern Japan Earthquake. Therefore, we used the names of facilities or places in eastern Japan as cues for machine learning. When preparing for disasters anywhere, we must collect the names of a facilities or places all over the world.

In this paper, we focused on tweets written in Japanese. In our future work, we will translate cue phrases from Japanese into other languages, and apply our method to tweets written in various languages.

REFERENCES

[1] S. Verma, S. Vieweg, W. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram, and K.M. Anderson, "Natural Language Processing to the Rescue? Extracting Situational Awareness Tweets During Mass Emergency," Proc. the Fifth International AAAI Conference on Weblogs and SocialMedia, Barcelona, Spain, pp. 385–392, July 2011.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors," Proc. 19th International World Wide Web Conference, NC, USA, April 2010.

[3] I. Shizu Miyamae Iwanaga, T. M. Nguyen, T. Kawamura, H. Nakagawa, Y. Tahara, and A. Ohsuga, "Building an Earthquake Evacuation Ontology from Twitter," Proc. 2011 IEEE International Conference on Granular Computing, Kaohsiung, Taiwan, November 2011.

[4] H. L. Truong, L. Juszczyk, A. Manzoor, and S. Dustdar, "Escape - an Adaptive Framework for Managing and Providing Context Information in Emergency Situations," Proc. 2nd European Conference on Smart Sensing and Context, vol. 4793, pp. 207–222, Lake District, UK, October 2007.

[5] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety Information Mining — What can NLP do in a disaster —," Proc. 5th International Joint Conference on Natural Language Processing, pp. 965-973, Chiang Mai, Thailand, November 2011.

[6] D. Davidov, "Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns," Proc. 2009 Conference on Empirical Methods in Natural Language Processing, pp. 267-275, Singapore, August 2009.

[7] A. Ishino, H. Nanba, and T. Takezawa, "Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries," Proc. 18th International Conference on Information Technology and Travel & Tourism, Innsbruck, Austria, January 2011.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. 18th Conference on Machine Learning, pp. 282-289, MA, USA, June 2001.

[9] http://mecab.sourceforge.ne

[10] C. Corinna, and V. Vladimir, "Support-Vector Networks," J. Machine Learning, vol.20, No.3, pp. 273-297, 1995.

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