Local Information Extraction Technology Twitter Natural Language

Technology Reports

Technology to Discover Local Events Using Twitter

Twitter is a social networking service that enables real-time, large-scale sharing of information on a wide range of topics. This article describes technology that automatically discovers information from Twitter for local events being held throughout Japan using natural language processing technology, and introduces a "town event information" service enabled by this technology.

Research Laboratories *Wataru Yamada*
Service Innovation Department *Keiichi Ochiai*
Haruka Kikuchi

1. Introduction

With the popularization of smartphones, there are now many types of location information services available. For example, Trip Advisor®*1 is a service that provides information and reviews on hotels and restaurants all over the world, while the Gurutabi®*2 service provides fine dining information for various localities in Japan. To make these location information services more attractive, it's important to provide the latest information about sightseeing spots, events, and specialties in localities.

However, it takes a lot of effort to maintain provision of up-to-date information for localities. Moreover, frequent updating is also required for information on local events to ensure the freshness of information because new events are held often, and handling these manually has its limitations.

To solve this issue, we have developed technology to automatically discover information about events being held in localities in Japan using Twitter*3. Twitter is a social networking service that enables users to post and share text messages that are up to 140 characters long (called tweets*4). Twitter enables users to share large amounts of information on various topics such as things happening in their own lives, new products, news and events. In particular, Twitter makes it easy for anyone to announce an event, not only highly public events such as fireworks displays or local festivals, but also a wide range of other events such as store fairs or indie band performances.

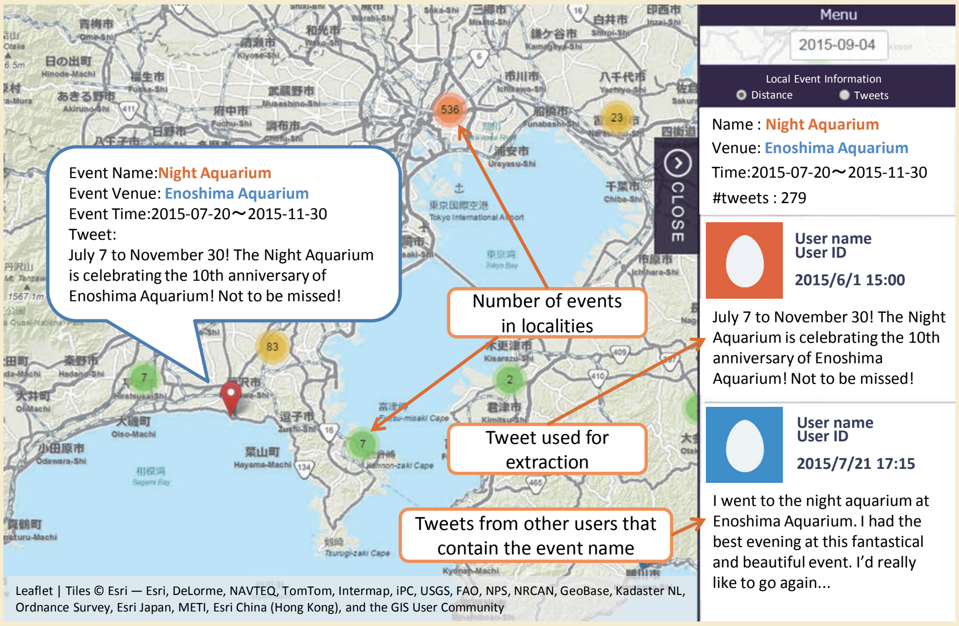
The technology we have developed uses natural language processing technology*5 to automatically discover event information from tweets. Not only does the technology determine whether or not there is an event, but also can extract the event name, location, and time with approximately an 80% to 90 % degree of accuracy.

This article describes this automatic local event information discovery technology, and also describes a "town event information" service using this technology.

2. Overview of the Local Event Discovery Technology

The following describes an overview of the local event discovery technology. **Figure 1** shows the operations screen of the demonstration application. This application was prepared to visualize and demonstrate this technology, and displays the automatically extracted event

*1 **Trip Advisor®:** An international service that provides information on restaurants and hotels, and reviews. The trademark and registered trademark of TripAdvisor, LLC.



Event Name:Night Aquarium
Event Venue: Enoshima Aquarium
Event Time:2015-07-20～2015-11-30
Tweet:
July 7 to November 30! The Night Aquarium is celebrating the 10th anniversary of Enoshima Aquarium! Not to be missed!

Menu
2015-09-04
Local Event Information
○ Distance ● Tweets

Name : Night Aquarium
Venue: Enoshima Aquarium
Time:2015-07-20～2015-11-30
#tweets : 279

User name
User ID
2015/6/1 15:00
July 7 to November 30! The Night Aquarium is celebrating the 10th anniversary of Enoshima Aquarium! Not to be missed!

User name
User ID
2015/7/21 17:15
I went to the night aquarium at Enoshima Aquarium. I had the best evening at this fantastical and beautiful event. I'd really like to go again...

Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

CLOSE
536
23
83
7
2
7
Number of events in localities
Tweet used for extraction
Tweets from other users that contain the event name

* These tweets are fictional examples..

Figure 1 Image of event discovery technology prototype

information as the time of the event and its location, which makes it easy for users to find event information in various localities. From automatically extracted local event information, Figure 1 shows events assumed to be held on Friday September 4, 2015. Information for each event contains the three pieces of information, the event name, location, and time, and also contains tweet information from which these pieces of information were extracted. For example, the balloon in Figure 1 shows three pieces of information - the event name as "Night Aquarium," the location as "Enoshima Aquarium," and the time as "July 20, 2015 to November 30, 2015." Also, users can search through tweets that contain the event name to see other users' tweets about the event. The numbers on the map in Figure 1 show the total number of events in various localities automatically extracted in a similar manner to the "Night Aquarium" event. For example, on September 4, there are 536 events being held in the Tokyo area.

The actual number of events discovered changes with seasons, with holidays and weekends, and with business days, although in general, the system can extract around 150 new events nationwide every day, or around 4,000 to 5,000 per month. This is the largest event database in the country.

Conventionally, events were extracted mainly by determining locations from a sharp rise in posts of tweets appended with location information [1] [2]. Because it's presumed that the number of tweets from around the location of the event will change in this way, this method extracts information for large-scale happenings where many tweets are posted such as earthquakes or performances by famous artists, but cannot extract small scale events that have relatively lower numbers of related tweets. Furthermore, this method cannot extract information prior to an event.

With our event discovery technology, even if there are not many tweets with location information attached, it's possible to extract event information using machine learning*6 technology that focuses on the natural language characteristics of the tweets that contain event information. Hence, even a single tweet announcing an event can be used to extract information, which makes it possible to discover not only large-scale events

*2 **Gurutabi®:** A service that introduces local specialties throughout Japan. A registered trade mark of Gurunavi, Inc.

*3 **Twitter:** A registered trademark of Twitter, Inc. in the United States and other countries.

*4 **Tweet:** A term used to refer to postings on the micro-blogging service provided by Twitter, Inc.

*5 **Natural language processing technology:** Technology to process the language ordinarily used by humans (natural language) on a computer.

*6 **Machine learning:** A mechanism allowing a computer to learn the relationship between inputs and outputs, through statistical processing of example data.

but also small-scale events held locally.

3. System Structure and Process Flow

The local event discovery technology consists of a location name extraction section and an event information extraction section, as shown in **Figure 2**. Processing for these sections is as follows.

3.1 Location Name Extraction Section

Using tweets in Japanese, the location name extraction section analyses which location the tweet is discussing, and then associates the tweet with the location. This association takes three steps - (1) to (3) shown in Figure 2. Firstly, morphological analysis[*7] is performed on the Japanese language tweets (Figure 2 (1)). Next, a list of location names that includes such information as locality and facility names and flags describing ambiguity (described later) is referenced, and tweets containing nouns that match location names are extracted (Figure 2 (2)). Finally, filtering is performed for tweets that contain ambiguous location names (Figure 2 (3)). Location names that are ambiguous include names that are also used for people, or names that are used in multiple locations that cannot be uniquely determined, for example, the surname “Matsushima” and the sightseeing locality “Matsushima” in Miyagi Prefecture. An example of a location name used in more than one place is “Maruyama Park” in Kyoto and “Maruyama Park” in Hokkaido. To eliminate these ambiguities, processing with co-occurrence words and machine learning is used so that filtering is performed

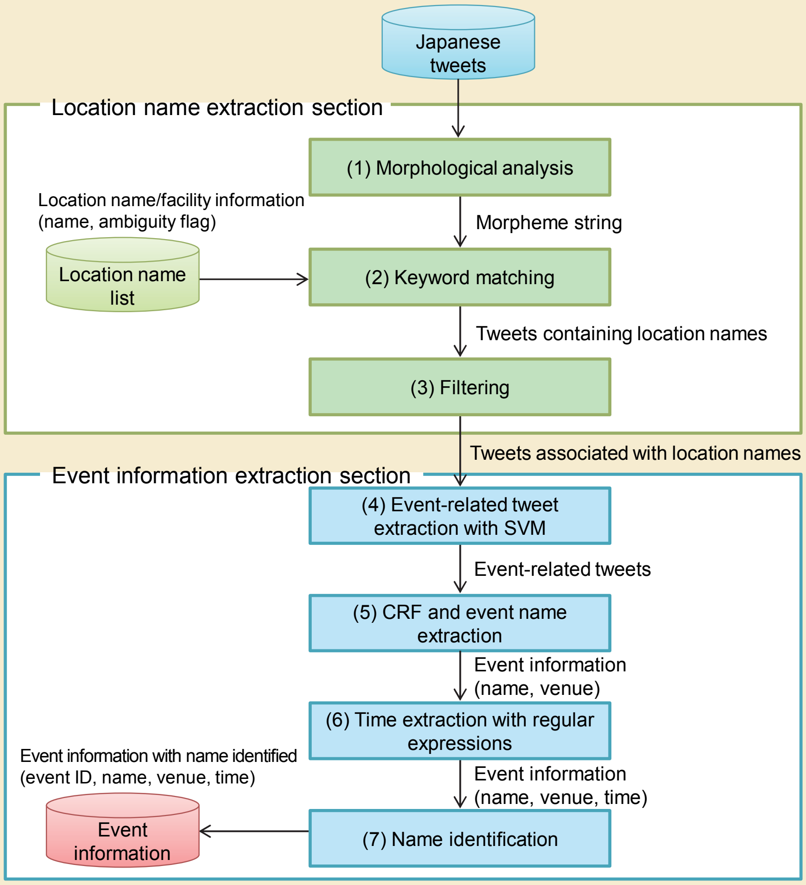

Japanese tweets
Location name extraction section
(1) Morphological analysis
Location name/facility information (name, ambiguity flag)
Morpheme string
Location name list
(2) Keyword matching
Tweets containing location names
(3) Filtering
Tweets associated with location names
Event information extraction section
(4) Event-related tweet extraction with SVM
Event-related tweets
(5) CRF and event name extraction
Event information (name, venue)
(6) Time extraction with regular expressions
Event information with name identified (event ID, name, venue, time)
Event information (name, venue, time)
Event information
(7) Name identification

Figure 2 Overview of entire processing for local event discovery technology

[*7] **Morphological analysis:** Technology that separates the smallest parts of words in text that have meaning, called morphemes.

correctly. See reference [3] for details of this filtering.

After the above processing, the location name extraction section outputs the Japanese language tweets associated with the location name to the event information extraction section, with ambiguity removed.

3.2 Event Information Extraction Section

The event information extraction section extracts the event name and time from tweets associated with location names. In general, this process consists of four steps. Firstly, event-related tweet extraction processing is performed to extract tweets from the tweets associated with location names that contain information about the event name and time (Figure 2 (4)). Secondly, event names are extracted from tweets that contain event information. Also, the location name associated with tweets used in the location name extraction section is assigned to event venue information (Figure 2 (5)). Thirdly, the event time is extracted from the tweets from which the event name was extracted (Figure 2 (6)). Fourthly, for the extracted event information, using degrees of similarity with the venue and event name, name identification processing is performed to determine whether the names are the same, then each piece of the event information is given an ID (Figure 2 (7)). The details of these processes are described below.

1) Event-related Tweet Extraction

Using machine learning, event-related tweet extraction only extracts tweets related to events from tweets already associated with location names. In this process, a classifier[*8] learns characteristics of words that appear often in event-related tweets such as "hold" or "festival," and then event-related tweets are extracted from tweets associated with location names. To process large numbers of tweets quickly, an algorithm called linear Support Vector Machine (SVM) [4] is used. This algorithm is split into two phases, as shown in **Figure 3**.

(a) Firstly in the learning phase, tweets are collected, and each tweet is visually checked to see whether or not it is an event-related tweet, and either an "event-related" or "event-unrelated" la-

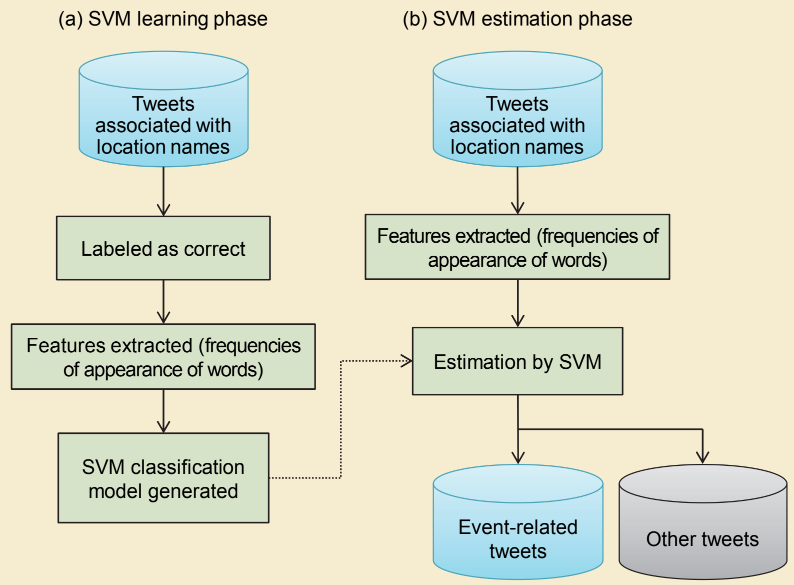

(a) SVM learning phase
(b) SVM estimation phase
Tweets associated with location names
Labeled as correct
Features extracted (frequencies of appearance of words)
SVM classification model generated
Tweets associated with location names
Features extracted (frequencies of appearance of words)
Estimation by SVM
Event-related tweets
Other tweets

Figure 3 Event-related tweet extraction with SVM

[*8] **Classifier:** A device that classifies inputs into a number of predetermined classifications based on their feature values.

bel is attached to each tweet. Next, features[*9] are extracted from labeled tweets. The number of times words appear is used for these features. For this reason, if there are tweets with words such as “hold” or “exhibition” that appear often in event-related tweets but not in non-event related tweets, these tweets are judged to be event-related tweets.

(b) In the estimation phase, the classifiers configured in the learning phase are used to determine whether tweets associated with location names contain information about events. Tweets which are determined to contain event information are output to event name extraction processing.

2) Event Name Extraction

Event name extraction uses SVM to extract event names from tweets judged to be event related. This process uses a machine learning algorithm called Conditional Random Fields (CRF)[*10] [5] to extract event names.

Event name extraction by CRF is also divided into two phases, as shown in **Figure 4**.

(a) In the learning phase, the sections that contain the event names and section unrelated to the event name in the event-related tweet are labeled. Then, CRF learns various characteristics such as words like “festival” that appear often in expressions and character strings such as “October fest” that appear often as event names, and then generates a classification model. Word readings and writings, parts of speech and number of characters are used as features.

(b) Using the classification model generated in the learning phase, the estimation phase determines which parts of tweets judged to be event related are related to event names. Then, tweets for which event names have been extracted are output for event time extraction.

3) Event Time Extraction

(1) Extraction using regular expressions

Regular expressions are used to extract event time from tweets. The use of regular expressions is one method of natural language processing which determines whether character strings match predetermined

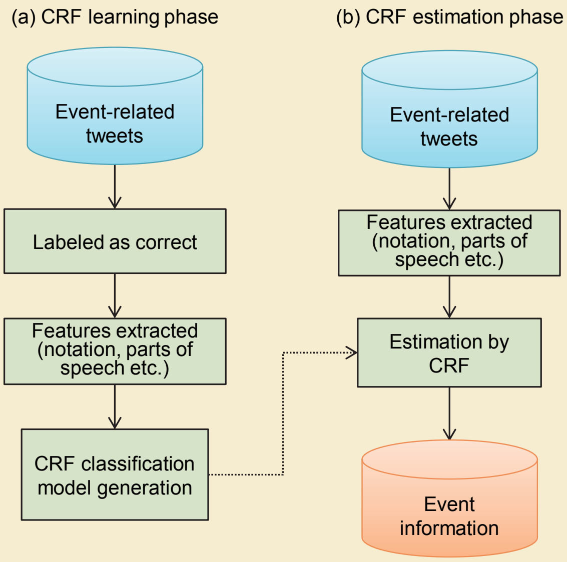


Figure 4　Event name extraction by CRF

[*9] **Feature:** The feature values in natural language processing.

[*10] **CRF:** A type of method for assigning pre-defined labels to a sequence of input entities based on the feature values of the entities.

patterns and makes it possible to extract those matching character strings from text. For example, by defining the pattern "¥d{2,4}/¥d{1,2}/¥d{1,2}" the character stings "2016/1/1" or "16/12/31" can both be extracted, since both 2 and 4 digits are defined for "year" and 1 and 2 digits are defined for month and day etc. Firstly, many date-related patterns are registered in advance for event time extraction, and then dates are extracted from text by matching those patterns.

(2) Date supplementation

Extracted dates do not necessarily contain all "year/month/day" date information, for example dates such as "1/1" or "Today." For such dates, the date that the tweet was posted is referenced and used to supplement the date so it can be read as year/month/day. Also, to handle dates that are not single dates such as "from January 1 to 3, 2016" that depict a period, the text is checked for indications of a period such as 'from,' 'to' or "-" between extracted dates. Thus, if there are indications of a period, information surrounding dates etc. is checked to extract the period for the event.

The above processes extract the three pieces of information about events - the event name and time, and the event venue assigned by the location name extraction section.

4) Name Identification Processing

Although event information up to the event time is extracted with this series of processes, there can be duplicate extractions due to the fact the multiple tweets can be posted informing of the same event.

Also, because different users might use different text characters or spelling to indicate an event name, such as "Future of the 21st Century Exhibition" or "21st Century Future Exhibition," the same events can be contained in event information with different names.

To solve this, this process uses the two pieces of event name and venue information to determine whether events are the same and then assigns the event an ID. These are numbers used for managing event information-the same ID is assigned to duplicated event information.

As shown in **Figure 5**, this process groups extracted event information by event venues. Next, all possible pairs are created from events with the same venues. Then, the event names for the events with the same venues are computed for degree of similarity, and determined to be the same event if this degree is above a certain threshold. These events names are then assigned the same event ID. The longest common subsequence ratio is used for the degree of similarity of event names. As shown in Figure 5, this process enables summarization of events with similar names that are to be held at the same venue.

4. The "Town Event Information" Service

Since May 12, 2015, a "town event information" service has been available in the DOCOMO d-menu real-time search corner. As shown in **Figure 6**, using this technology, this service enables users to display information to find topical events close to their location from event information collected for their current locality. Users can also search for information for events other than those in their locality by specifying locations and dates with maps and calendars, which is useful when traveling or going out.

5. Conclusion

This article has described an overview of technology to automatically discover local event information from tweets, and has introduced the "town event information" service provided by using this technology. In future, using this technology, we plan to produce content with nationwide event information to create services that can be used for local development. We will also research and develop local information extraction technologies for information other than events, such as information on local specialties and related reviews.

References

[1] L. Chen and A. Roy: "Event Detection from Flickr Data through Wavelet-based Spatial Analysis," Proc. of the 18th ACM Conference on Information and Knowledge Management, 2009.

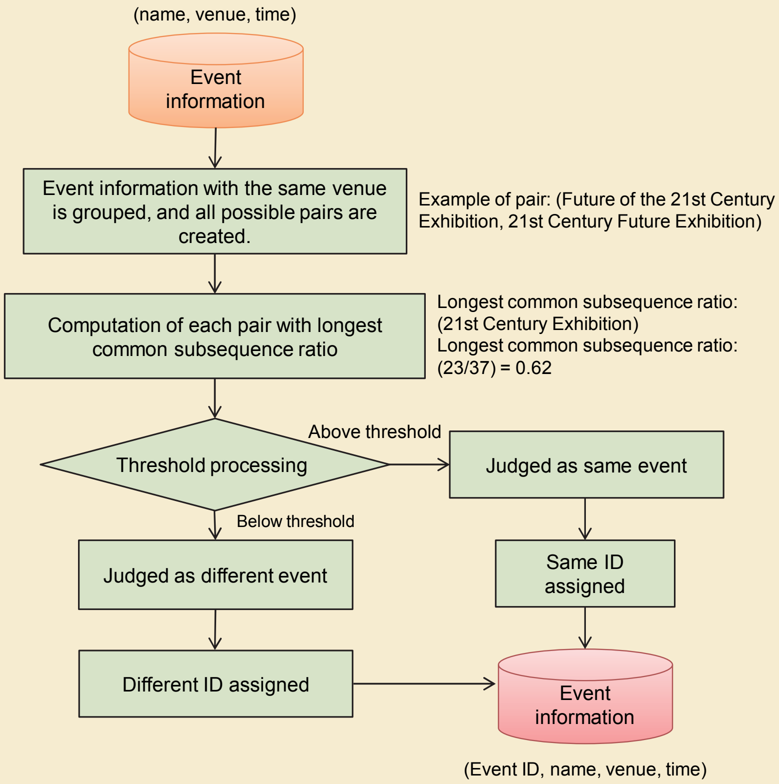
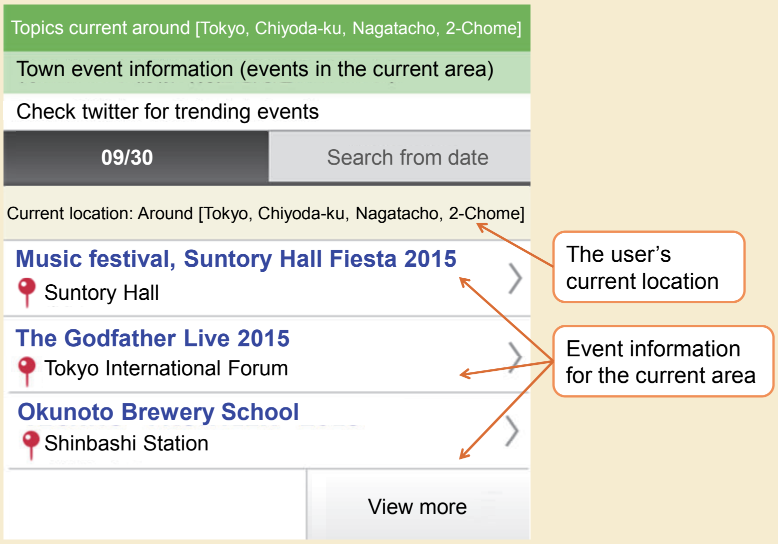

(name, venue, time)
Event information
Event information with the same venue is grouped, and all possible pairs are created.
Example of pair: (Future of the 21st Century Exhibition, 21st Century Future Exhibition)
Computation of each pair with longest common subsequence ratio
Longest common subsequence ratio: (21st Century Exhibition)
Longest common subsequence ratio: (23/37) = 0.62
Threshold processing
Above threshold
Judged as same event
Below threshold
Judged as different event
Same ID assigned
Different ID assigned
Event information
(Event ID, name, venue, time)

Figure 5　Name identification of event information


Topics current around [Tokyo, Chiyoda-ku, Nagatacho, 2-Chome]

Town event information (events in the current area)

Check twitter for trending events

Current location: Around [Tokyo, Chiyoda-ku, Nagatacho, 2-Chome]

Music festival, Suntory Hall Fiesta 2015
📍 Suntory Hall

The Godfather Live 2015
📍 Tokyo International Forum

Okunoto Brewery School
📍 Shinbashi Station

The user's current location

Event information for the current area

*This is provided only in Japanese at present.

Figure 6　Image of “Town event information” service screen

[2] R. Lee and K. Sumiya: "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection," Proc. of 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp.1-10, 2010.

[3] K. Ochiai et al.: "Analysis of Location-related Tweets and its Applications," NTT DOCOMO Technical Journal, Vol.16, No.2, pp.29-35, Oct. 2014.

[4] Corinna. C and Vladimir. V: "Support-Vector Networks, Machine Learning," Vol.20, pp.273-297, 1995.

[5] L. John, M. Andrew and C. Feramdo: "Conditional randomfields: Probablistic models for segmenting and labeling sequence data," Proc. of the Eighteen International Conference on Machine Learning, pp.282-289, 2001.