International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:11, 2008

Extraction of Temporal Relation by the Creation of Historical Natural Disaster Archive

Suguru Yoshioka, Seiichi Tani and Seinosuke Toda

Abstract**—** In historical science and social science, the influence of natural disaster upon society is a matter of great interest. In recent years, some archives are made through many hands for natural disasters, however it is inefficiency and waste. So, we suppose a computer system to create a historical natural disaster archive. As the target of this analysis, we consider newspaper articles. The news articles are considered to be typical examples that prescribe the temporal relations of affairs for natural disaster. In order to do this analysis, we identify the occurrences in newspaper articles by some index entries, considering the affairs which are specific to natural disasters, and show the temporal relation between natural disasters. We designed and implemented the automatic system of "extraction of the occurrences of natural disaster" and "temporal relation table for natural disaster."

Keywords**—** Database, digital library, corpus, historical natural disaster, temporal relation

I. Introduction

IN recent years, researches for the "Digital Archives" are widely applied in various disciplines [1], [3], [4], [6], [11], [16]. This "archive" represents the resources which are researched in the various disciplines, such as ancient documents, old graphical contents, moving images and so on. By digitization of these resources, transmission of the information on the Internet, it makes more people are familiar with the various disciplines. In particular, digital archive system is widely researched as data accumulation, information transmission, discovery science and problem solving. That is, for the development and spread of information technology and computer network, we need to note the effective usage of analysis resources and we expect to discover an advanced utilization. And, we have implemented a digital archive system to utilized many resources in our university[1].

In this paper, we take particular note of natural disaster, make the historical natural disaster archive by extracting information of an occurrence of an earthquake from a newspaper article. Moreover, we define the necessary items for historical science as the index entries of our archive. Then, in this paper, we pay our attention to the extraction technique in natural language processing, develop a textual archive. We show each natural disaster is *situated* [2] by the indexes of each news article in our archive. We also show our archive is utilized effectively for analyses of (i) community-based disaster frequency, (ii) a temporal relation between natural disasters and (iii) a social impact, by using our archive and index information.

In the following section, we present our project for digital archives and the importance of historical natural disaster. In Section III, we explain the computer system to make the historical natural disaster archive based on natural language processing, and show some examples and the results of our system. In Section IV, we explain the temporal structure in the archive, define some constraints and rules, propose the system to analyze a temporal relation between natural disasters. In the final section, we discuss some branching points of our theory and summarize our contribution.

II. Digital Archive for Natural Disaster

In this section, we explain the digital archive system in our university. We show the feature of natural disasters and the attempt for the historical natural disaster archive.

A. Digital Archive Project

In our university, there have been the project to make digital archives. We have built a computer system for digital archives. Our system consists mainly of database storage and retrieval. The researchers of the various academic disciplines put their resources into our database system for digital storage, such as ancient documents, and old paintings, digitalized map, electroencephalographic recordings and so on. We have provided these digitalized resources to general public through our Internet-based system[2]. This system allows general public to look up for the objects he or she is interested. So, we need to study the technique of digitalization and the usage of digitalized resources. This attempt is widely applied in the field of information sciences.

B. Archives for Historical Natural Disaster

In historical science and social science, influences of natural disasters on society and human beings are a matter of great interest. Thus far, many historians and sociologists have proposed the data creation of the time-line of natural disaster history. In particular, for earthquake, it has been classified depending on temporal sequence, earthquake intensity, location and so on. However, for analyzing the relation between earthquake and social world, we require a comprehensive study. That is, we need to note (i) population density in the disaster area, (ii) temporal relation between a certain

S. Yoshioka, S. Tani and S. Toda are with the institute of information science, Nihon University.
e-mail:s-yoshio@chs.nihon-u.ac.jp, sei-ichi@tani.cs.chs.nihon-u.ac.jp, toda@cssa.chs.nihon-u.ac.jp

[1]"Academic Frontier" Project for Private Universities: matching fund subsidy from Ministry of Education, Culture, Sports, Science and Technology, 2002-2006

[2]See Internet site http://da.chs.nihon-u.ac.jp/da/

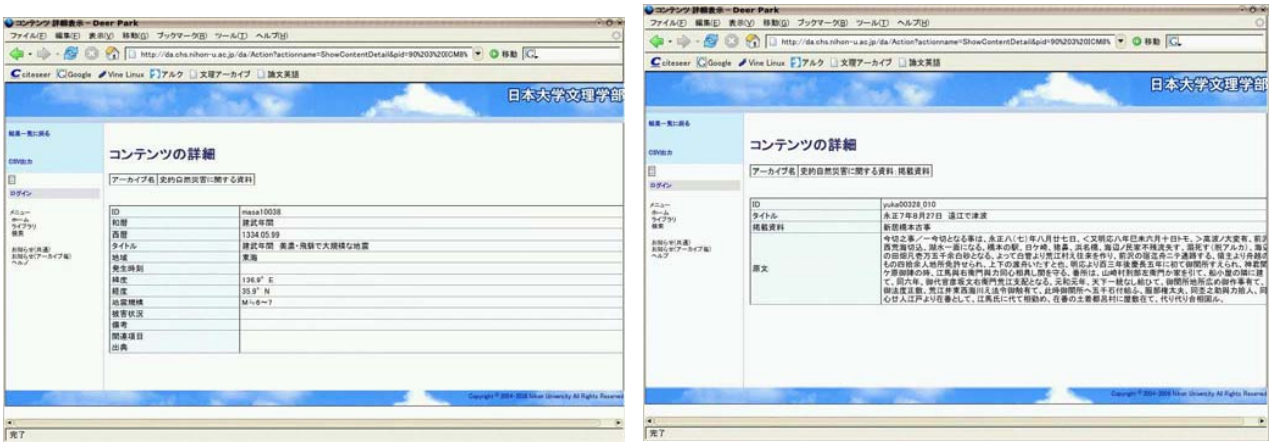

Fig. 1. Archives for natural disasters

earthquake and other earthquakes, (iii) relation between a certain earthquake and other natural disasters.

As a part of the project, our university historians have attempted to create two historical archives. One of the archives is a "information about natural disaster" and the other is a "information of the documents about natural disaster." The "information about natural disaster" consists of the tuple ⟨ ID, Japanese calender, dominical year, title, region, date, latitude, longitude, earthquake size, disaster damage, remarks, pertinent details, source ⟩ as shown in the left-hand side of Fig. 1. The "information of the documents about natural disaster" consists of the tuple ⟨ ID, title, document name, original text ⟩ as shown in the right-hand side of Fig. 1. where "title" represents a kind of disasters.

The descriptions about natural disaster have a difference between contents depending on references, i.e. original text, even though these descriptions represent the same disaster. So, we need to make the "information of the documents about natural disaster" to examine the social impact. By the "information about natural disaster" and the "information of the documents about natural disaster," we can classify natural disasters by date and location, and reference, respectively.

These two archives are made through many hands, hence there is inefficiency and waste. Moreover, they have related the "information about natural disaster" to the "information of the documents about natural disaster," however they have not relate a certain natural disaster to other disasters in each archive. For example, there is a big difference between "There were an earthquake" and "There were an earthquake while the biggest typhoon hit" for the social impact. That is, we need to design and implement the automatic generation system for the natural disaster archive. And then we expect to analyze temporal and causal relation between natural disasters.

III. Computer System to Make the Natural Disaster Archive

We propose a system based on the natural language processing to prepare the historical natural disaster archive.

A. Index Information

The objective of this paper is to analyze the temporal structure of a sequence of occurrences. As the target of this analysis, we consider newspaper articles for a natural disaster. We used Mainichi-Shinbun corpus [3] which is a textual corpus.

We extracted the news stories which include the noun "earthquake" from this corpus. Our system was implemented in Perl (Ver.5.8.2) on Vine Linux 5.2 CR. Then, the number of all news articles was 339489 items and the number of all news articles with "earthquake" was 5320 items [4]. Here, we define some index entries as follows:

Definition 1 (Index entry): The natural disaster archive consists of the following index entries.

news date, prefecture, region,
date of a disaster, time of a disaster
Richter scale, earthquake intensity,
temblor's epicenter, body of a news article,

where "news date" is the date when the news item is printed. "prefecture" and "region" are place where an earthquake hit. In particular, "region" denotes the traditional classification in Japan, Japan divides into 9 areas (Hokkaido, Tohoku, Kanto, Chubu, Kinki, Chugoku, Shikoku, Kyushu and Ryuku areas.) And, "date of disaster" and "time of disaster" are day and time that earthquake hit, respectively. "Richter scale" and "earthquake intensity" are the scale of an earthquake and the Japanese scale of an earthquake announced by Japan Meteorological Agency, respectively.

These index entries are seen as the important items of earthquake research by historians. So, we extracted the above information from news articles if those information are said in newspaper articles. Here, we define some symbols and operators for index entries.

[3]Mainichi-Shinbun is one of major newspaper company in Japan. Our research progress is provided by using the Mainichi-Shinbun corpus for 1998 year, 2000 year and 2001 year.

[4]We counted the newspaper articles on a per-page basis in corpus

TABLE I
AN EXAMPLE OF THE CO-OCCURRING NOUNS WITH "EARTHQUAKE"

7243	earthquake	989	observation	⋮	⋮
1890	earthquake intensity	915	emergence	28	water stoppage
1555	seismic source	⋮	⋮	⋮	⋮
1494	it	176	collapse	19	landslide

TABLE II
AN EXAMPLE OF THE CO-OCCURRING VERBS WITH "EARTHQUAKE"

1055	said	55	collapsed	⋮	⋮
857	say	55	continued	25	be destroyed
799	be	55	released	⋮	⋮
⋮	⋮	⋮	⋮	9	be burnt down

Definition 2 (Signature): The following symbols and operators are used.

NI	a set of news articles,
$ndate(n)$	news date,
$pref(n)$	Japanese prefecture,
$reg(n)$	Japanese region,
$ddate(n)$	date of disaster,
$dtime(n)$	time of disaster,
$Rich(n)$	Richter scale,
$intens(n)$	earthquake intensity,
$epic(n)$	temblor's epicenter,
$\vee$, $\wedge$	logical connectives,

where $n \in NI$, and also we use $n_1, n_2, \cdots$ for news articles. And, parentheses and punctuation are added if necessary.

Example 1: Given the following news article n_1:

98'.01.16 : １６日午前１０時５８分ごろ、関東地方を中心に地震があった。気象庁によると、震源地は千葉県南部で、震源の深さは約６０キロ、マグニチュードは４・８と推定される。

Jan. 16th, 1998 : On 16th at 10.58AM, An earthquake hit mainly in Kanto area. The Japan Meteorological Agency said that the temblor's epicenter was southern part of Chiba prefecture, the depth of hypocenter was 60 kilometer and the quake measured 4.8 on the Richter scale.

Then, our system extracts Jan. 16th 1998, Chiba prefecture, Kanto area, 16, 10.58AM, 4.8, southern part of Chiba prefecture as $ndate(n_1)$, $pref(n_1)$, $reg(n_1)$, $ddate(n_1)$, $dtime(n_1)$, $Rich(n_1)$ and $epic(n_1)$, respectively.

Here, the tuple $\langle\, pref(n_1), reg(n_1), ddate(n_1), dtime(n_1), Rich(n_1), intens(n_1), epic(n_1) \,\rangle$ denotes about the "information about natural disaster" as shown in Subsection II-B. And, the tuple $\langle\, ddate(n_1), n_1 \,\rangle$ denotes about the "information of the documents about natural disaster."

B. Feature of Japanese Language and Newspaper Corpus

It is hard to construe in Japanese because there is no morpheme boundary. And, Japanese have many lexical ambiguities. So, we extracted nouns and verbs by using the Japanese language morphological analysis system JUMAN [9] and the Japanese language syntax analysis system KNP [10]. Additionally, we treated a compound noun as one noun. And also, we treated a verb with auxiliary verb as one verb to identify an occurrence and information of the tense. We show the portion of the extracted nouns and verbs in Table I and Table II, respectively.

Where the numbers in Table I and Table II represent the frequency of appearance. As shown in Table I and Table II, the co-occurrence frequency is not relevant to the causal relation with earthquake. So, we define the constraint rules to represent a temporal structure in the following section.

We show the feature of corpus and extraction technique. Mainichi-Shinbun corpus is a textual corpus which consists of *ID*, *news title*, *date of news*, *running text of article*, *a result of the morphological analysis* and so on[5]. In this corpus, one paragraph denotes a newspaper article per page. We extracted index entries for each one paragraph of an article as follows:

Rule 1 (Extraction): We extract the index entries by the following procedures.

n $\in$ **NI** : Extract the *running text of article*
ndate(n) : Extract the *date of news* in corpus
pref(n) : We make the **meta-data** such as a tree structure for municipality depending on the result of the analysis by JUMAN and KNP. Then, we extract the prefecture if there is a description of town, city or prefecture by using the meta-data.
region(n) : If there is a description of region, we extract it, otherwise if $pref(n)$, extract a region by using the meta-data.
ddate(n) : In Japanese, the word "日" is necessary to describe the date. It means about the "th" which is a postfix of the date in English. So, we extract the numerical values that precede "日" as $ddate(n)$.
dtime(n) : If there is the descriptions "時" and "分" which mean "hour" and "minute", respectively,

[5]See Internet site www.nichigai.co.jp/sales/mainichi/mainichi-data.html

exstract the numerical values that precede these words.
Rich(n) : If there is the description "マグニチュード" or "M" that preced the numerical value, extract this numerical value as $Rich(n)$.
intens(n) : If there is the description "震度" that preceds the numerical value, exstract this value as $intens(n)$.
epic(n) : If there is the descriptions "震源地は" and "で (or punctuation)", extract the description between these words as temblor's epicenter, where "震源地は" and "で" mean "temblor's epicenter is (or was)" and "in (or on)" in English, respectively.

Our system is only for Japanese Mainichi-Shinbun corpus because of the morphlogical analysis, but the core algorithm can be easily applied to other languages.

C. Constraints by Index Information

In Subsection III-A, we took particular note of the word "earthquake", exstracted some news articles from the corpus. However, in exstracted news stories, there are many stories which have little or nothing to do with the occurrence of earthquake. The following example shows an unconcerned news article of the occurrence of earthquake.

Example 2:
Note the folowing news article n:

地震保険めぐる裁判　契約より約款が優先　火災保険訴訟に影響もーー神戸地裁
In a trial about an earthquark insurance at District Court in Kobe, it is sentenced that adhesive terms and conditions are given priority over contract. So, it has no small effect on fire insurance lawsuit.

Because the purpose of this research is to compile the fact of the occurrence of earthquake, we have to except unconcerned news articles from our archive. In Mainichi-Shinbun corpus, a lot of news articles for the occurrence of earthquake have a distinct description for the *Richter scale* or the *earthquake intensity*. And a lot of news articles which does not say about a region have a compound noun such as Earthquake Research Committee, Research Center for Earthquake Prediction, a metropolitan epicentral earthquake and so on. That is, we define the following rule.

Rule 2 (Constraint by region and earthquake size): $\forall n \in NI$, if $reg(n) \wedge (Rich(n) \vee intens(n))$ exists, then extract n. Otherwise eliminate.

We represent a set of news articles by Rule 2 as NI'. Here, I show the results of experiments. We checked the number of articles with $reg(n)$, $Rich(n)$ and $intens(n)$ for the newspaper articles of 1998 year, 2000 year and 2001 year. We show a result of the experiment as follows:

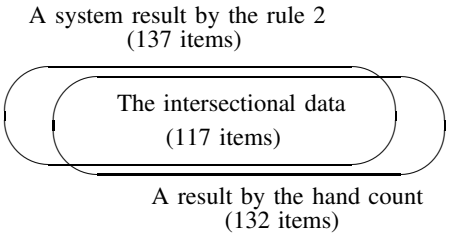

Fig. 2. Recall and Precision

No. of all articles	339489 items
No. of articles about an earthquake	5320 items
No. of $reg(n)$	3439 items
No. of $Rich(n)$	986 items
No. of $intens(n)$	1307 items
No. of $Rich(n) \vee intens(n)$	1780 items
No. of $reg(n) \wedge (Rich(n) \vee intens(n))$	1531 items

In particular, we experimented with the creation of the natural disaster archive for news articles from January 1998 to June 1998 as a preparatory analysis. In this experiment, the number of all news articles and resultant news articles are as follows:

No. of all articles	61637 items
No. of articles about an earthquake	581 items
No. of resultant articles by Rule 2	137 items

Thus, by Rule 2, garbage news articles are 444 items, however there are correct 15 items in these garbage stories. So, the ratio is 96.6% that our system was able to eliminate garbage articles, but 3.4% that our system eliminated the correct articles. And, we checked a correct data by the hand count and as a consequence, there are 132 items. We show the relation between the result of our system and the correct data by the hand count in Fig. 2.

Thus, by Rule 2, *recall* and *presicion* are obtained as follows:

$$\text{Recall} \quad \frac{117}{132} = 88.6\%$$

$$\text{Presicion} \quad \frac{117}{137} = 84.1\%$$

Where recall is the number of answers the system got right divided by the number of possible right answers. Precision is the number of answers the system got right divided by the number of answers the system gave. It is generally accepted that there is a trade-off relation between recall and presicion.

Here, we show some examples for the failure of Rule 2 as follows:

Example 3:
Given the following news articles:

1998.4.15 箕面などで地震
April 15th, 1998 *There was an earthquake in Minoh city.*

1998.4.30 岩手山で、火山性地震
April 30th, 1998 *There was a volcanic earthquake on Mount Iwate.*

The news article of straightforward expression is failed to extract by Rule 2. Many news articles say abridged information about earthquake in the morning paper but detailed information about the earthquake in the evening paper. And, the news articles about volcanic earthquake or swarm earthquakes prioritize the continuity and the notice of earthquake over the magnitude of earthquake. So, those articles are not selected correctly by Rule 2, even though the articles say the occurrence of earthquake.

Furthermore, we focus attension of a difference between the "news date" and the "date of disaster" as shown in Definition 1. In the nature of newspaper, news article that was written in day-month-year format is a rare case, because many news articles were written on the very day or the following day of an earthquake. So, only date is exstracted as $ddate(n)$. Then, there is an ambiguity between $ndate(n)$ and $ddate(n)$. We show an example as follow:

Example 4:
Given the following news article n:

> *2000.01.13 :* １７日で阪神大震災から丸５年が経過するのを前に、政府は１３日、首都圏直下型の大地震発生を想定した災害訓練を実施した。
> *January 13th, 2000 : Before 5 years have elapsed since the Great Hanshin Earthquake struck on 17th, the Japanese government conducted disaster drills on 13th to get ready for surprises when the metropolitan epicentral earthquake comes.*

For the above example, the $ndate(n)$ and the $ddate(n)$ are 13th and 17th, respectively. Then, the $ddate(n)$ produces a temporal ambiguity, i.e. we can not assess whether the 17th represents the future or the past. However we can assume the $ddate(n)$ represents the past time, because news article about an occurrence of earthquake was generally written in the fact that an earthquake struck. So, we define the following rule.

Rule 3 (Constraint by the date): $\forall n \in NI'$, If $ndate(n) < ddate(n)$ then eliminate n.

We represent a set of news articles after application of Rule 3 as NI''. Here, we show a result of the experiment. We created a natural disaster archive for the news articles of 1998 year, 2000 year and 2001 year by Definition 1 and Rule 2. After that, as a result of the application of Rule 3, 23 news articles were exstracted and eliminated. All of these news articles were unconcerned about a fact that an earthquake struck. We expect to increase precision by Rule 3.

IV. Temporal Structure in the Natural Disaster Archive

The natural disaster archive which created in Section III consists of a set of tuples $\langle\ ndate(n),\ pref(n),\ reg(n),\ ddate(n),\ dtime(n),\ Rich(n),\ intens(n),\ epic(n),\ n\ \rangle$. But there is an ambiguity for the temporal structure between tuples. For example, swarm earthquakes strike from day to day, however our system considers such a series of earthquakes as a set of different earthquakes for the simply by virtue of having a different news date. So, we define the following rule to distinguish between unexpected earthquake and a series of earthquakes.

Rule 4 (Constraint by the news date): $\forall n_1, n_2 \in NI''$, $ndate(n_2) < ndate(n_1)$, if $3 < (ndate(n_1) - ndate(n_2))$ then n_1 and n_2 are unconcerned disaster to each other.

And also, two earthquakes have no temporal causal relation, if these two news paper articles have different region no matter what $ndate(n)$ and $ddate(n)$ of these articles are same. So, we define the following rule.

Rule 5 (Constraint by the region): $\forall n_1, n_2 \in NI''$, if $reg(n_1) \neq reg(n_2)$ then n_1 and n_2 are unconcerned disaster to each other.

Here, we show the archive after the application of Rule 4 and Rule 5 in Fig. 3. In Fig. 3, each row represents one news article, viz. an occurrence of an earthquake. And each empty row includes a horizontal line represents a result of the sorting by Rule 4 or Rule 5. However, as shown in Fig. 3, while there is a series of earthquakes, if an earthquake hits in a different area, the information about a series of earthquakes drops out from archive. Because empty rows are inserted in the archive whenever there is an overlapping in time but not in space and vice versa. So, we made several archives on the regional basis, applied Rule 4 and the following rule.

Rule 6 (Constraint by the prefecture): $\forall n_1, n_2 \in NI''$, if $pref(n_1) \neq pref(n_2)$ then n_1 and n_2 are unconcerned disaster to each other.

Here, we show a result of the experiment. We experimented with the same resource as shown in Subsection III-C. As a result, we could divide into a series of earthquakes correctly in the 7 regions of the 9 regions. And, 2 regions which remain as an issue to be solved include quack-prone area and several prefectures with isolated island. We show a result of the experiment as follows:

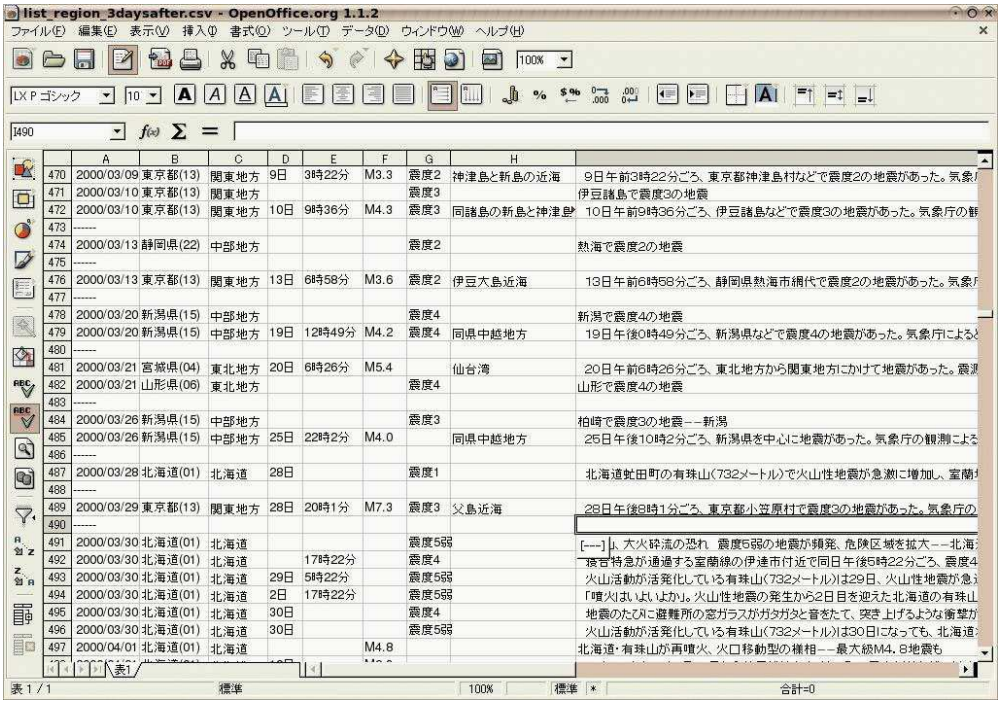

Fig. 3. The temporal structured archive for the natural disaster

Regions	The number of articles
All	1090 items
Hokkaido area	42 items
Tohoku area	84 items
Kanto area	485 items
Chubu area	169 items
Kinki area	164 items
Chugoku area	109 items
Shikoku area	22 items
Kyushu area	13 items
Ryuku area	2 items

For the newspaper articles of Kanto area and Chubu area, we could not categorize the earthquakes correctly.

V. Conclusion

We summarize our contribution as follows. (i) We proposed the index entries so as to make the narural disaster archive from the news paper corpus. (ii) In addition, giving some constraints as rules, we could distinguish a series of earthquakes from the natural disaster archive. (iii) Lastly, as a result, we could analyze a temporal relation of natural disasters.

And, there are several future subjects: (i) we need to note the linguistical structure and aspect for increasing the coverage of recall and precision. (ii) In addition, we need to analyze natural disasters except an earthquake, such as volcano, seismic surges, tropical cyclone, and so on.

References

[1] R. B. Allen and J. Schalow. : Metadata and data structure for the historical newspaper digital library, In CIKM '99: *Proceedings of the 8th Intrernational Conference on Information and Knowledge Management*, pp. 147-153, New York, USA, 2005. ACM Press.

[2] J. Barwise and J. Seligman. : *Information Flow*, Cambridge University Press, 1997.

[3] K. Bontcheva, D. Maynard, H. Cunningham and H. Saggion. : Using human language technology for automatic annotation and indexing of digital library content. In *ECDL '02 Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 513-625, London, UK, Springer-Verlang, 2002.

[4] G. Crane and A. Jones. : The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection, *Proceeding of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 31-40, 2006.

[5] M. Deegan, E. Steinvel, and E. King. : Digitizing historic newspapers : progress and prospects, *RLG Diginews*, 6(4), August, 2002

[6] M. Federico, D. Giordani and P. Coletti : Development and evaluation of an Italian broadcast news corpus, *Proceeding of the LREC*, v2, pp. 921-924, 2000.

[7] R. Goldblatt : *Logics of Time and Comuputation*, Second Edition, CSLI Lecture Note No.7, Center for the Study of Language and Information, Stanford University, 1992.

[8] H. Kamp. : *Some remarks on the logic of change. Part I*, In Roherer (ed.), *Time, Tense and Quantifiers*, Niemeyar, Tübingen, 1979.

[9] S. Kurohashi and M. Nagao : Japanese morphological analysis system JUMAN version 3.61 manual. In Japanese, 1999.

[10] S. Kurohashi and M. Nagao : Japanese syntax analysis system KNP manual. In Japanese, 1994.

[11] M. Markkula and E. Sormunen : End-user searching challenges indexing practices in the digital newspaper photo archive, *Information Retrieval*, 1, pp.259-285, 2000.

[12] M. Moens. and M. Steedman : Temporal Ontology and Temporal Reference, *Computatinal Linguistics*, 14(2), pp.15-28, 1988.

[13] Y. Shoham : *Reasoning about Change*, The MIT Press, 1988.

[14] S. Yoshioka, K. Kaneiwa and S. Tojo : Occurrence Logic with Temporal Heredity, *Proceeding of the IICAI-03*, pp.1296-1309, 2003

[15] S. Yoshioka and S. Tojo : Many-dimensional Modal Logic of Tense and Temporal Interval and its Decidability, *Journal of the Japanese Society for Artificial Intelligence*, Vol. 21, No. 3, pp. 257-265, 2006.(in Japanese)

[16] I. H. Witten, K. J. Don, M. Dewsnip and V. Tablan. : Text mining in a digital library. *Int. Journal on Digital Libraries*, 4(1): 56-59, 2004.