Sophisticated Text Mining System for Extracting and Visualizing Numerical and Named Entity Information from a Large Number of Documents

Masaki Murata, Tamotsu Shirado, Kentaro Torisawa
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{murata, shirado, torisawa}@nict.go.jp

Masakazu Iwatate
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan masakazu-i@is.naist.jp

Koji Ichii
Hiroshima University
1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527, Japan
ichiikoji@hiroshima-u.ac.jp

Qing Ma
Ryukoku University
Otsu, Shiga 520-2194, Japan
qma@math.ryukoku.ac.jp

Toshiyuki Kanamaru
Kyoto University
Yoshida-Nihonmatsu-cho, Sakyo, Kyoto 606-8501, Japan
kanamaru@hi.h.kyoto-u.ac.jp

Abstract

We have developed a system that can semiautomatically extract numerical and named entity sets from a large number of Japanese documents and can create various kinds of tables and graphs. In our experiments, our system semiautomatically created approximately 300 kinds of graphs and tables at precisions of 0.2–0.8 with only 2 h of manual preparation from a 2-year stack of newspapers articles. Note that these newspaper articles contained a large quantity of data, and all of them could not be read or checked manually in such a short amount of time. From this perspective, we concluded that our system is useful and convenient for extracting information from a large number of documents. We have constructed a demonstration system. In this paper, we briefly describe the demonstration system.

Keywords: *Text Mining System, Visualization, Numerical Information, Named Entity, Graph*

1 Introduction

Text documents contain many kinds of numerical and named entity (NE) information, such as temperature, humidity, and places. Extracting such information and preparing graphs from it can prove useful for extracting information from text documents [16, 1, 3]. MuST workshop has been held for helping researches extracting numerical information from text documents and preparing graphs from it [3]. We have constructed a system that can semiautomatically extract numerical and NE sets from a large number of Japanese documents and can make various kinds of tables and graphs, such as a graph where the vertical axis can indicate the wind speed for a typhoon, the horizontal axis can indicate the central atmospheric pressure, and a tag for each plot can indicate the place where the typhoon appeared [10]. The system can list the various kinds of numerical and NE information contained in a large number of documents using a small amount of human labor. It is also useful for knowing what kinds of information exist in these documents.

There have been some related studies. Matsushita et al. prepared a graph from information stored in a database [7], and Namba et al. and Murata et al. extracted trend information from documents and prepared graphs, plotting the temporal information on the horizontal axis and numerical values on the vertical axis [12, 9]. Murata et al. extracted numerical pairs from a document set concerning a certain topic and prepared scatter charts from them [8]. Swan et

al. made a visualized overview indicating when major topics occur in a large number of newspaper documents [17], but they did not extract detailed information on each topic. Shinyama et al. extracted related NE information from a large number of documents [15], but they did not extract numerical information or graphs. There are no studies that discuss the semi-automatic formulation of various kinds of graphs from a large number of text documents, including various kinds of numerical and NE information.

Our system can extract various kinds of graphs from a large number of text documents, including different types of information. It enables users to understand what kind of numerical and NE information is included in such documents and visualize this information using graphs. It can also plot a graph that includes three or more kinds of numerical information. Graphs are easy to comprehend and they facilitate the understanding of information in documents.

In Section 2, we explain our system. In Section 3, we describe experiments using numerical information only. In Section 4, we describe experiments including NE information. In Section 5, we describe experiments using word category dictionary. In Section 6, we describe our demonstration system.

2 System

Our system consists of the following three components.

Component 1—"Component for creating a list of key expression sets": A large number of documents are first inputted into the system. The system creates and outputs a list containing sets of key expressions that will be used to extract and merge numerical and NE sets. Key expressions are classified into three categories: item units, kinds of named entities (NE kinds), and item expressions. We used person, location, organization, and artifact names, as well as time, date, money, and percentage expressions as the NEs. This procedure was similar to that used in IREX [14]. The system extracts item units, item expressions, and NEs. We used YamCha [4], a chunker with a support vector machine [2], for extracting NEs. We used words, parts of speeches, characters, and so on as the features for machine learning. We used the morphological analyzer ChaSen [6] to extract item units and item expressions. The system extracts a sequence of nouns adjacent to numerical values as the item units and extracts a sequence of nouns as the item expressions. It then creates sets of item units, kinds of NEs, and item expressions appearing many times in the same sentence in documents, and outputs a list of sets to the user in the order of the frequency in which the set appears in the documents. For example, the system extracts a set consisting of the item expression "starting time", item unit 1: "°C", item unit 2: "%", item unit 3: "m/s", kind of NE 1: "location name", and kind of NE 2: "organization name" because these expressions appear together frequently in the same sentence in the documents.

Table 1. Selection of key expressions in data using two item units

Item expression	Item units		Freq.
sakunen (last year)	*sai* (age)	*nin* (number of people)	189
kakaku (price)	*en* (yen)	*heihoumeetoru* (square meters)	188

Component 2—"Component for allowing the user to select sets of key expressions": A user selects sets of key expressions from the list that he or she judges to be useful. We show an example of the list in Table 1. The user judges that the key expressions on the first line are unrestricted, and that a graph made from these expressions will include several topics and therefore will not yield a coherent graph. Hence, the user does not select this set. The user judges that the key expressions on the second line are restricted and that a graph made from these expressions will yield a coherent graph with regard to land prices. Hence, the user selects this set.

Component 3—"Component for creating graphs for selected sets of key expressions": For each selected set, the system identifies locations in the sentences where the item units, kinds of NEs, and item expressions (a set of key expressions) appear near each another. The system then extracts a set of key expressions described in the sentences as a numerical set; it also extracts numerical values appearing with item units and uses the connection between the numerical values and item units as numerical expressions. It then extracts an NE of the same kind as the NE of a set of key expressions. For example, for the sentence "The weather conditions for Kyoto at the starting time are temperature of 14°C, humidity of 62%, and wind velocity of 2 m/s", when a set of key expressions is given as the item expression: "starting time", item unit 1: "°C", item unit 2: "%", item unit 3: "m/s", and kind of NE 1: "location name", the system extracts a set consisting of the item expression: "starting time", numerical expression 1: "14°C", numerical expression 2: "62%", numerical expression 3: "2 m/s", and kind of NE 1: "location name: Kyoto" from this sentence. The system then gathers the extracted numerical sets and NEs and uses them to create tables or graphs. We used Excel to create the graphs. We manually inputted the headings for the tables and axes for the graphs.

Table 2. Number of sets of key expressions

No. of unit items	2	3	4	5	6	7
Total	511343	80345	23071	19210	50125	32647
> 4	28648	4174	1287	372	91	11
Checked	3000	1411	1287	372	91	11
Selection	60	35	20	0	0	0

Table 3. Evaluated results

No. of unit items	Eval. A	Eval. B	Ave. num. of plot
2	0.47 (28/60)	0.72 (43/60)	36
3	0.37 (13/35)	0.71 (25/35)	14
4	0.70 (14/20)	0.85 (17/20)	4
Total	0.48 (55/115)	0.74 (85/115)	24

3 Experiments using numerical information only

3.1 Experiments

We conducted experiments to make various kinds of graphs from a large number of documents by using our system. In these experiments, we used a 2-year stack of Mainichi newspaper articles written in Japanese from 1998 and 1999 [5] (220,078 articles). We used one item expression and 2–7 item units as the key expressions. Our experimental results are shown in Table 2. In this table, the first line ("No. of unit items") indicates the kind of key expression set and indicates the number of unit items. We always used one item expression for the key expressions. "Total" indicates the total number of extracted sets of key expressions. (We extracted key expressions appearing in the same sentence at least once as a set of key expressions.) Further, "> 4" indicates the number of extracted sets of key expressions appearing in the same sentence at least five times. "Checked" indicates the number of sets of key expressions checked by a subject. The subject checked a list of sets of key expressions from the top of the list and checked the sets of key expressions whose numbers equal the value described in the "Checked" line. During this checking process, the subject judges whether or not each set of key expressions can be used for extracting sets of numerical values and making a graph. "Selection" indicates the number of sets of key expressions judged to be useful by the subject.

Next, our system created graphs by using the sets of key expressions selected by the subject. We evaluated the created graphs. The results are shown in Table 3. The first column "No. of unit items" indicates the kind of sets of key expressions and the number of unit items. "Eval. A" and "Eval. B" are abbreviations for "Evaluation A" and "Evaluation B". Evaluation A indicates that a graph in which 75% or more of the points were correct was judged to be correct and Evaluation B indicates that a graph in which 50% or more of the points were correct was judged to be correct. Here, the points related to a certain topic and extracted as the correct value from the documents were judged to be correct. We used Evaluations A and B because the graphs where 75% or 50% of the plots are correct would be useful for recognizing the outline of the data set and facilitate the manual modification of incorrect plots. In Evaluation B, our system obtained accuracy rates of approximately 0.7–0.8. The system created 55 graphs satisfying Evaluation A and 85 graphs satisfying Evaluation B. Table 3 also shows the average number of plots in the graphs created by our system. It indicates that the graphs corresponding to two unit items have many plots and graphs corresponding to more unit items have fewer plots.

The experiments involved human labor of about 1 h for checking the sets of key expressions. That is, our system semiautomatically created approximately 100 types of graphs from a 2-year stack of newspaper articles with human labor of about 1 h. It should be noted that a two-year stack of newspaper articles contains a large quantity of data and it cannot be read or checked by human beings within a short time. From this perspective, we concluded that our system is very useful and convenient.

We examined the error cases and found that the main reason that errors occurred was that many things were related to the same item units, so the numerical sets related to different things were extracted. For example, *en* (yen) was extracted as an item. However, yen was related to many things such as the "amount of sales", "net profit", "monthly amount", and "annual amount". The system extracted mixed numerical sets including these quantities and could not produce a coherent graph.

3.2 Graphs created by the system

In this section, we show some graphs created by our system.

A graph created using two unit items is shown in Figure 1. Figure 1 was created by using *chuushin kiatsu* (central atmospheric pressure) as the item expression and *hekuto pasukaru* (hectopascal) and *meetoru*

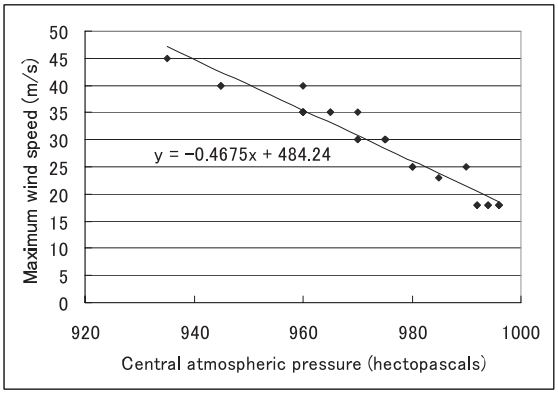

Figure 1. Graph using two unit items for typhoon

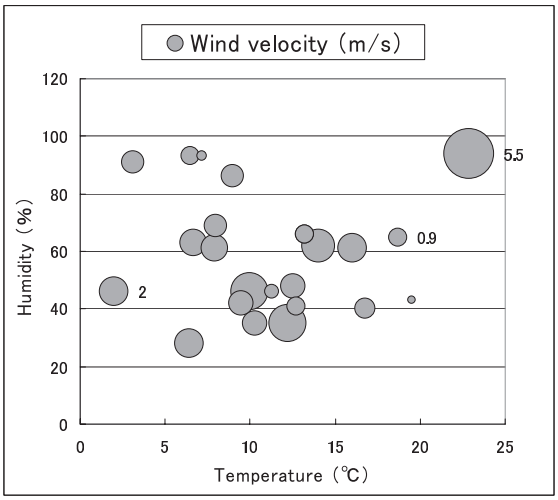

Figure 2. Graph using three unit items for marathon

(m/s) as the unit items. This graph can be considered to be related to typhoons. In this graph, the horizontal axis indicates the central atmospheric pressure of a typhoon and the vertical axis indicates the maximum wind speed of a typhoon. From this graph, it can be seen that when the pressure is lower, the wind speed is higher. Further, different speeds can be encountered for the same pressure. We calculated a single regression line for the points and plotted it in the figure. The equation for the line can be used to estimate the maximum wind speed from the central atmospheric pressure.

A graph created using three unit items is shown in Figure 2. Figure 2 was created using *staato ji* (at the starting time) as the item expression and °C, "%", and *meetoru* (m/s) as the unit items. The documents used for making the graph described the topic of the

Table 6. Evaluation results using NE information

No. of NEs	Evaluation A	Evaluation B	Ave num. of plot
1	0.67 (2/3)	0.67 (2/3)	1165
2	0.50 (1/2)	1.00 (2/2)	2165
3	0.00 (0/2)	0.50 (1/2)	363
4	0.40 (2/5)	0.60 (3/5)	254
5	0.21 (12/58)	0.69 (40/58)	64
6	0.19 (13/69)	0.83 (57/69)	14
7	0.00 (0/1)	1.00 (1/1)	6
Total	0.21 (30/140)	0.76 (106/140)	103

Table 7. Evaluation results using numerical and NE information

No. of NEs	Evaluation A	Evaluation B	Ave num. of plot
1	0.75 (15/20)	0.85 (17/20)	249
2	0.14 (1/ 7)	0.29 (2/ 7)	76
3	0.56 (14/25)	0.76 (19/25)	99
4	0.70 (31/44)	0.93 (41/44)	105
5	0.73 (30/41)	0.83 (34/41)	25
6	0.62 (33/53)	0.79 (42/53)	8
Total	0.44 (124/190)	0.56 (155/190)	74

marathons. The graph can be considered to be related to marathons. In this graph, the horizontal axis indicates the temperature, the vertical axis indicates the humidity, and the diameter of each circle indicates the wind velocity at the staring time of the marathon. This graph reveals the air condition during each marathon. For example, the point near the upper-right corner reveals high temperature (23°C), high humidity (94%), and high wind velocity (5.5 m/s).

Our system could semiautomatically extract numerical information and create such interesting graphs.

4 Experiments including NE information

4.1 Experiments

We also carried out experiments on NE information. In these experiments, we used a 2-year stack of accrued Mainichi newspaper articles from 1998 and 1999 [5] (220,078 articles). We performed these experiments using one item expression and 1–8 NE kinds as the key expressions and the experiments using one item expression, 1–8 NE kinds, and two item units as the key expressions. Our experimental results are sum-

Table 4. Number of sets of key expressions

	Use of NE information only			Use of numerical and NE information		
Number of NEs	Total	> 4	Selection	Total	> 4	Selection
1	1000156	107007	3	422672	5961	20
2	972428	82780	2	494159	5182	7
3	606420	44183	3	434067	5562	25
4	271353	18749	6	255301	4615	44
5	66561	3466	69	87232	1928	41
6	8929	288	75	18727	104	53
7	325	1	1	748	1	0
8	41	0	0	123	0	0

Table 5. An example of a list using one NE and two item units

Item expression	Item unit		NE	Example of NE	Freq.
meijinsen (championship game)	*kyoku* (game)	*ki* (period)	Person	Koji Tanigawa, Yasumitsu Sato, Yoshiharu Habu, Makoto Nakahara, Toshiyuki Moriuchi, Taku Morishita, Tadahisa Maruyama, Hihumi Kato, Keita Inoue, Akira Shima	514

marized in Table 4. In this table, the first line ("No. of NEs") indicates the kind of key expression set, representing the number of NEs. "Total" indicates the total number of extracted sets of key expressions. (We extracted the key expressions appearing in the same sentence at least once as a set of key expressions.) Further, "> 4" indicates the number of extracted sets of key expressions appearing in the same sentence at least five times.

We extracted the top 100 sets among the key expressions appearing at least five times and manually checked them. "Selection" indicates the number of sets of key expressions that have been judged to be useful by the subject. It is difficult to judge each key expression to be useful or not by only observing the NE kinds in the experiments involving the use of NE information. Therefore, when a key expression was manually checked in the experiments that used NE information, the top 10 NEs were provided for each NE kind for consultation. An example of a list used in the checking process is shown in Table 5.

Next, we evaluated the data obtained by the selected key expressions. The results are shown in Tables 6 and 7. In Tables 6 and 7, Evaluation A (Eval. A) and Evaluation B (Eval. B) indicate a graph and table where 75% and 50% or more of the points or data items were judged to be correct, respectively. Here, the points in the graph or data items in the table were related to a certain topic and those points or data items extracted as correct values from the documents were judged to be correct. By using our system, we obtained accuracy rates between 0.2 and 0.5 in Evaluation A and that between 0.5 and 0.8 in Evaluation B using our system. The system created 30 tables and 124 graphs satisfying Evaluation A and 106 tables and 155 graphs satisfying Evaluation B. Tables 6 and 7 also list the average number of plots or data items in the graphs and tables created by our system.

Table 8. Data extracted using NE information only

(a) Player throwing slider and his team

Player	Team
Inoue	JAL
Brosse	Yakuruto
Yano	Takanabe
Crossford	Seibu
Sakai	Kintetsu
Yoshii	Mets
...	....

(b) Missile and its country

Missile	Country
Shaheen	Pakistan
Rodong-1	North Korea
Taep'o-dong 2	North Korea
Arrow	Israel
Taep'o-dong	North Korea

The experiments involved manually checking the sets of key expressions for 1 h. That is, our system semiautomatically created approximately 200 kinds of graphs from newspaper articles accumulated over a 2-year period in a process that only involved manual preparation for about 1 h. Note that these newspaper articles contained a large quantity of data and could not be read or checked manually in such a short time. From these results, we concluded that our system is useful and convenient.

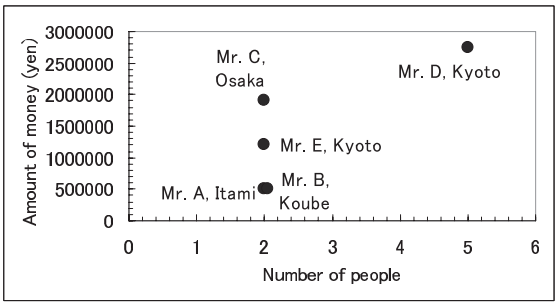

Figure 3. Graph using numerical and NE information for bribery charges

4.2 Graphs created by the system

Some examples of data extracted using our system are shown in Table 8 and Figure 3. Table 8 uses NEs and an item expression. Table 8(a) uses an item expression of "suraidaa" (slider) and two NEs of the person's and organization's names as the key expressions. We can identify the players throwing sliders and their team from the table. Table 8(b) uses the item expression of "dandou misairu" (ballistic missile) and the two NEs of artifact and location name as the key expressions. From the table, we can identify the names of ballistic missiles and the countries that have them. Other than the abovementioned data, we could obtain various kinds of data such as "igo" and "shougi" data indicating the date, place, organizer, and game players, as well as domiciliary search data indicating the searched organization, date, location, person, money, and related laws.

Figure 3 shows the data obtained when NEs and item expressions were used as the key expressions. The item expression "shuuwai zai" (bribery charge), item units "nin" (number of people), and "en" (yen) and the NE kinds of person and location names are used as the key expressions. The vertical axis in the figure indicates the number of people related to bribery charges, and the horizontal axis indicates the amount of money related to these bribery charges. The person and location names are displayed as labels for each plot. Although the system could obtain the persons ' names, we used anonymous substitutes here. Other than the data, we could obtain various kinds of graphs indicating which floor the room on fire was located, the number of stories in the building, the name of the room's occupant, and the time at which the fire began. We could obtain various kinds of graphs indicating the order of an athletic game, its length, the players' names, the organizations they belonged to, and the location of the game.

Table 9. Modified BGH category numbers

Semantic marker	Original code	Modified code
Animal	[1–3]56	511
Human	12[0–4]	52[0–4]
Organization	[1–3]2[5–8]	53[5–8]
Products	[1–3]4[0–9]	61[0–9]
Parts of a living thing	[1–3]57	621
Plant	[1–3]55	631
Nature	[1–3]52	641
Location	[1–3]17	657
Quantity	[1–3]19	711
Time	[1–3]16	811
Phenomenon	[1–3]5[01]	91[12]
Abstract relation	[1–3]1[0–58]	aa[0–58]
Human activity	[1–3]58, [1–3]3[0–8]	ab[0–9]

5 Use of word category dictionary

We performed additional information extraction using a word category dictionary to extract many more kinds of information other than numerical and NE information. We used the Japanese thesaurus, *Bunrui Goi Hyou* [13], as the word category dictionary.

In BGH, each word has *a category number*. In the electronic version of BGH, each word has a 10-digit category number that indicates 7 levels of the 'is-a' hierarchy. The top five levels are expressed by the first five digits, the sixth level is expressed by the next two digits, and the last level is expressed by the last three digits.

We used the categories shown in Table 9, developed in the paper by Murata et al. [11]. First, we convert the first three digits of the category number, such as that in Table 9. We classified the words into a category based on the first two digits of the modified category number and used the classified words. We used 13 categories in total.

Here, we used these 13 categories similar to that when we used 8 NE kinds in our system. We introduce some examples obtained by using the categories as follows.

We obtained the data on delayed and cancelled trains extracted by using one category of the dictionary, two item units, and one item expression as the key expression. The example is shown in Table 10. Table 10 is the result obtained when we used the category "Human activity" in Table 9 as one category of the dictionary, *hon* (unit for the number of trains) and *nin* (unit for the number of people) as two item units, and *eikyou* (effect) as an item expression. This data indicates what kind of cause affects how many trains and how many people. A cause of delayed and cancelled

Table 10. Data on delayed and cancelled trains extracted using the category dictionary

Cause	Number of trains	Number of people
jiko (accident)	64	34000
jiko (accident)	16	15000
traburu (trouble)	50	700
traburu (trouble)	11	7000
jokyo sagyou (clearing operation)	16	10000
kakunin (confirmation)	4	3000
suto (strike)	39000	900000
...	...	...

trains was extracted using the category of "Human activity". A cause of delayed and cancelled trains is not expressed as a NE but expressed as a common noun. To extract expressions such as a cause, it is necessary to use information on a common noun. In this paper, we used the category dictionary to handle a common noun.

In addition to the data, we obtained various kinds of data such as the data including the dead person's name, the date and time of death, the occupation and managerial position of the deceased, and the cause of death by the method using the category dictionary. The occupation and managerial position of the deceased and the cause of death are not expressed by a NE but by a common noun. Therefore, the category dictionary was useful for the extraction of such expressions.

6 Demonstration System

We constructed a demonstration system (Japanese language version). The system can extract text data from Web news, extract numerical and NE sets from the text data, and display the sets using a graph. We checked the gasoline prices using this system. Figures 4 and 5 show our demonstration system. Figure 4 shows the manner in which our system extracts numerical and NE sets from the text data. Figure 5 shows the manner in which our system makes a graph. Figure 6 shows a graph made using our system. In the figure, the vertical axis indicates a day ("日") when gasoline is sold and the horizontal axis indicates the gasoline price (yen, "円"). From the figure, we found

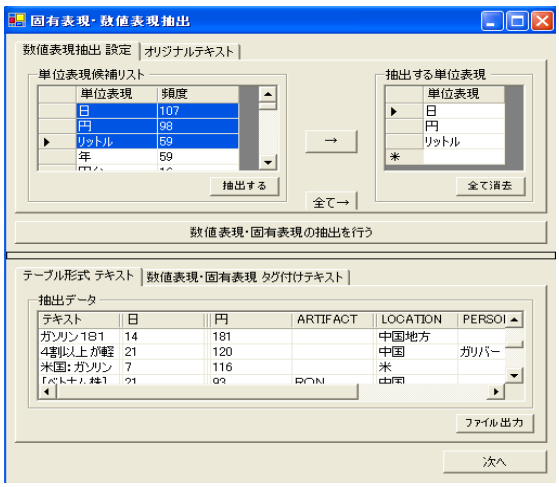

Figure 4. Extracting numerical and NE sets from the text data in our Japanese demonstration system

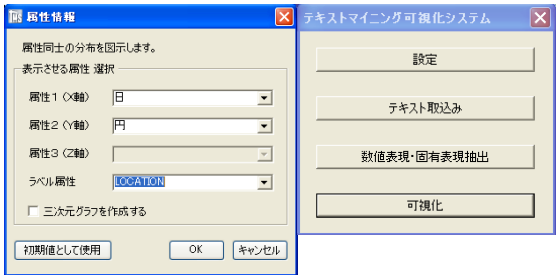

Figure 5. Making a graph in our Japanese demonstration system

that the gasoline price is comparatively low in USA ("米"), China ("中国"), and Vietnam ("ベトナム"). The gasoline price is comparatively high in Japan (e.g., Tsushima ("対馬"), Tyuugoku region ("中国地方"), Hokkaido ("北海道"), and Naha ("那覇")) and Korea ("韓国"). Extracting both numerical and NE information in our system played an important role for making this graph.

7 Conclusion

We constructed a system that could semiautomatically extract numerical and NE sets from a large number of documents and could yield various kinds of tables and graphs. To confirm the effectiveness of our system, we performed experiments using a two-year stack of newspaper articles. In these experiments, our system semiautomatically created approximately 300 kinds of graphs and tables with only 2 h of manual preparation. These newspaper articles contained a

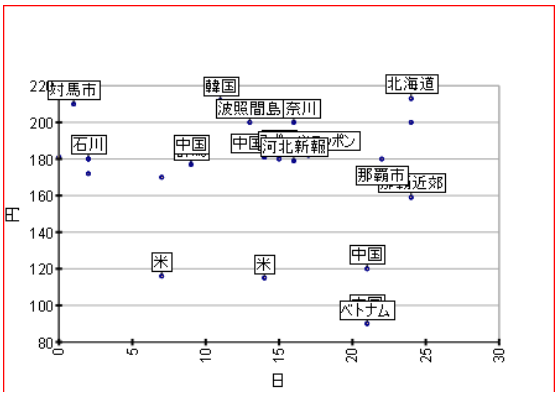

Figure 6. Graph made using our Japanese demonstration system

very large quantity of data, and all of them could not be read or checked manually in such a short time. Therefore, we concluded that our system is useful and convenient for extracting information from a large number of documents. Further, we constructed a demonstration system. In this paper, we briefly described the demonstration system.

Our system can be used for several applications. For example, our system can be used for an application system that outputs the information wanted by a user. When a user would like to know about weather or politics, the application system retrieves documents including the word "weather" or "politics" and extracts various kinds of numerical and NE information about weather or politics from the retrieved documents by using our system.

In the future, we would like to use text documents from the Web and make graphs from these documents. The main reason that our system had errors is that it sometimes extracted mixed numerical and NE sets related to more than one topic. We would like to use a clustering technique to divide mixed sets into coherent sets related to one topic.

References

[1] C. Chen. *Information Visualization, Beyond the Horison, Second Edition*. Springer, 2004.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[3] T. Kato, M. Matsushita, and N. Kando. MuST: A workshop on multimodal summarization for trend information. *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, 2005.

[4] T. Kudo. YamCha: Yet Another Multipurpose CHunk Annotator. http://www.chasen.org/ taku/ software/yamcha/, 2005.

[5] Mainichi Publishing. Mainichi Newspaper 1998-1999, 1999.

[6] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. 1999.

[7] M. Matsushita, H. Yonezawa, and T. Kato. Frame representation of user's requirement for automated data visualization. In *ECAI 2000*, volume 38, pages 631–635, 2000.

[8] M. Murata, K. Ichii, Q. Ma, T. Shirado, T. Kanamaru, S. Tsukawaki, and H. Isahara. Developing text mining and visualization system for numerical information pairs. *SCIS and ISIS*, 2006.

[9] M. Murata, K. Ichii, Q. Ma, T. Shirado, T. Kanamaru, S. Tsukawaki, and H. Isahara. Development of an automatic trend exploration system using the MuST data collection. *Proceedings of the ACL 2006 Workshop on Information Extraction Beyond The Document*, 2006.

[10] M. Murata, M. Iwatate, K. Ichii, Q. Ma, T. Shirado, T. Kanamaru, and K. Torisawa. Extraction and visualization of numerical and named entity information from a large number of documents. *2008 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2008)*, 2008. (to appear).

[11] M. Murata, K. Kanzaki, K. Uchimoto, Q. Ma, and H. Isahara. Meaning sort — three examples: dictionary construction, tagged corpus construction, and information presentation system —. *Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2001, Mexico City, February 2001 Proceedings*, pages 305–318, 2001.

[12] H. Nanba, N. Okuda, and M. Okumura. Extraction and visualization of trend information from newspaper articles and blogs. In *Proceedings of the 6th NTCIR Workshop*, pages 243–248, 2006.

[13] NLRI. *Bunrui Goi Hyou*. Shuuei Publishing, 1964.

[14] S. Sekine and H. Isahara. IREX project overview. *Proceedings of the IREX Workshop*, pages 7–12, 1999.

[15] Y. Shinyama and S. Sekine. Preemptive information extraction using unrestricted relation discovery. In *HLT-NAACL-2006*, pages 304–311, 2006.

[16] R. Spence. *Information Visualization*. ACM Press, 2001.

[17] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56, 2000.