Visualization for Statistical Term Network in Newspaper

Hideki Kawai Kazuo Kunieda Keiji Yamada
NEC C&C Innovation Research Laboratories
8916-47 Takayama-cho, Ikoma-city, Nara 630-0101, Japan
{h-kawai@ab, k-kunieda@ak, kg-yamada@cp}.jp.nec.com

Haruka Saito Masaaki Tsuchida Hironori Mizuguchi
NEC Common Platform Software Research Laboratories
8916-47 Takayama-cho, Ikoma-city, Nara 630-0101, Japan
{h-saitou@ia, m-tsuchida@cq, hironori@ab}.jp.nec.com

Abstract

In this paper, we propose a visualization method for global dynamics. Global dynamics of various events and statistics are important to analyze complex international issues such as environmental, economic and political problems. We have been developing a system which can extract a co-occurrence network of statistical terms based on a suffix pattern matching. However, the network structure consisting of thousands of statistical terms is too complicated to understand their causal relations briefly. So we propose a method for simplifying the network structure based on network complexity and language expressions. Our experimental result shows that a clique of the statistical terms corresponds to a certain topic or issue and causal relations can be described as a chain of the cliques on the network structure.

Keywords: *Global Dynamics, Co-occurrence Network, Statistical Terms*

1 Introduction

In our modern society, since various events and phenomena can interact each other, a local solution for a single problem may cause unexpected consequences. To solve complex problems such as issues about oil prices and global-scale climate change, we have to figure out global dynamics between various phenomena. Multimodal Summarization for Trend Information task (MuST)[5] is important because it can summarize and visualize trend information from various real world events.

Our research focus is on the extraction and visualization of a global dynamic network. An overview of a global dynamic network (Fig. 1) depicts that an event can affect other events in different domains. We exploit co-occurrence network of statistical terms[8, 7],

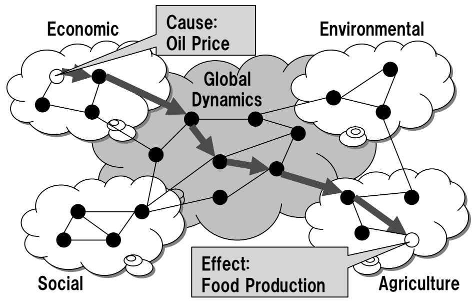

Figure 1. Global Dynamic Network

because statistical terms, like "birth rates," and "unemployment rates," have quantified values derived by measurment of events and phenomena in real world. We expect that observation of co-occurrence network of statistical terms can enhance our understanding of causal relations between various events. We have been developing an extraction method based on suffix patterns such as "率 (rates)" and "価格 (prices)," and also studying on functional dependencies between statistical terms.

However the structure of statistical term network is too complex to visualize. The reasons are two fold - (1) Network structure complexity and (2) Semantic structure complexity of statistical terms.

(1) Network structure complexity is derived from many nodes and edges in the network and also the existence of hubs co-occurring with too many other statistical terms. For instance, "人口 (population)" and "株価 (stock prices)" can be huge hubs because they are often written with other statistical terms in newspaper. These hubs can generate such a densely connected structure which is too complex to observe the relationship between nodes.

(2) Semantic structure complexity is derived from modifiers of statistical terms. For example, there are

same statistical terms with different modifiers, ”失業率 (unemployment rates)”, ”アメリカの失業率 (unemployment rates in America”, ”国内の失業率 (domestic unemployment rates)”, ”男性の失業率 (male unemployment rates)” and ”6 月の失業率 (unemployment rates in June).” In previous work, we classified these modifiers into four groups which represent the condition of statistics, i.e., object, subject, time span and region[8]. However, it is still unclear whether we can treat all these statistical terms as ”失業率 (unemployment rates)” or we have to treat them differently.

In this paper, we propose a simplification method of statistical term network based on both network and semantic structure complexity.

2 Related Works

Information Compilation is a fundamental technology that handles various information intelligently in order to enhance users' understanding of the information and to provide users with access to the information. MuST is one of tasks in this research area. Related works for building statistical term networks can be divided into two areas; extraction of statistical terms and finding the causal relations.

As the statistical terms extraction technique, Saito *et al.*[9] proposed a method using numerical expressions and their surrounding syntactic patterns which consists of certain word classes and particles. Fujihata *et al.*[1] utilized rules of dependency structure to extract numerical expressions and statistical terms. Mori *et al.*[6] defined statistical expression tags for a machine learning- based information extraction. While we exploit lightweight pattern matching focusing on common suffix of statistical terms.

For discovering causal relations, Sato *et al.*[10] used specific conjunctions which imply a causality such as ”〜したため (since)”, ”〜に伴って (along with)”. Another research group utilized a case frame dictionary to extract causal relations from sentences[11]. Inui *et al.*[3] reported that only 30% of causal relation in news articles are written explicitly with markers, but 70% of them are written implicitly without any markers. We focused on co-occurrence of statistical terms which can contain a wide range of relationships, including causal relations. In this paper, we also use a simplification technique for visualization of complex network structures.

3 Global Dynamics Visualization

This section describes the details of our global dynamics visualization. First, we describe a statistical term extraction technique for building a co-occurrence network. Next, we propose a network simplification method focusing on both network and semantic structure complexities.

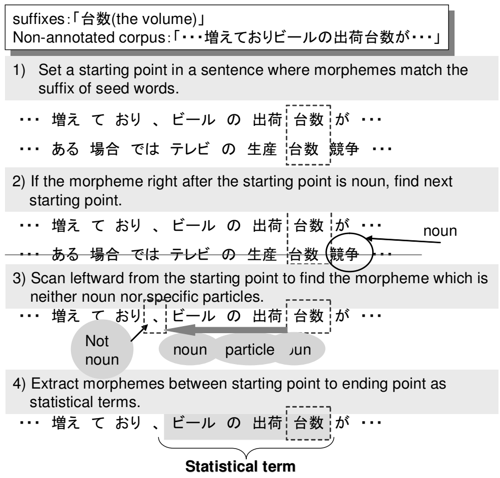

Figure 2. Statistical term extraction Algorithm

3.1 Extraction of Statistical Terms

Our statistical term extraction approach relies on lightweight pattern matching of common suffix of statistical terms[8]. We first prepare some statistical terms as seed words. Next, we split each seed word into several morphemes by using a morphological analyzer and we store the final one-to-three morphemes as suffix. Then, we extract noun phrases which have the same suffix as the seed words. This extraction process consists of four steps, as shown in Fig 2:

1. Set a starting point in a sentence where morphemes match the suffix of seed words.
2. If the morpheme right after the starting point is a noun, find another starting point.
3. Scan leftward from the starting point to find a morpheme which is neither a noun nor a specific particle.
4. Extract morphemes between starting point and ending point as statistical terms.

The above algorithm can extract modifiers which represent region, time span and conditions in which a statistic was computed. We assume that statistical terms consist of modifiers and base forms. We also defined four categories for modifiers: (a) object, (b) subject, (C) time span, and (d) region[8].

We defined the base form of a statistical term as the shortest sequence of morphemes having a statistically valid meaning. For instance, a base form of the statistical term, ”1999 年のアメリカの失業率 (U.S. unemployment rates in 1999),” corresponds to ”employment rates.” (a) Object is a modifier which represent measured objects such as people, organization and products. For example, in the statistical term "エアコンの出荷台数 (the volume of shipments of air-conditioners)," the object of measurement corresponds to "air-conditioners." (b) Subject is a modifier which represent those who conducted the statistical measurement. For example, in the statistical term "NEC のノートパソコンの出荷台数 (the volume of shipments of NEC's laptop PCs)," the subject corresponds to "NEC." (c) Time span is a modifier representing a measurement period. For example, "9 月の (in September)" and "1998 年上半期の (in the first half year of 1998)" fall into this category. (d) Region is a modifier representing a location where the statistic value was measured. For example, "アメリカの (in U.S.)" and "首都圏の (in the capital area)" fall into this category.

To analyze a causal relationship between statistical terms, we have to consider which modifier is actually important as the principle of causality. Consider the statistical term "アメリカの失業率 (U.S. unemployment rates)," some causal relationships are regionally specific and others may be universal. We can not decide which modifier is important based on only the modifiers themselves, but need to take into account the neighboring nodes in the statistical term network.

3.2 Simplification of Co-occurrence Network

After the extraction of statistical terms, we can construct a network by calculating co-occurrence between the statistical terms. In this paper, we defined "the co-occurrence" as the case when two statistical terms appear in the same paragraph in a news article. In the statistical term network, nodes correspond to statistical terms, and edges correspond to the co-occurrence of two terms at both ends.

The statistical term network is too dense and complex to visualize relationship between terms. Fig. 3 shows an example of a statistical term network. Fig. 3 shows only terms in two hops from the statistical term "リサイクル率 (recycle ratio)", but it is too busy to observe relationships between related terms. The main reason is some general statistical terms like "人口 (population)" can be hubs in the network, and make the network structure more complex by forming very dense clusters. So we propose a network simplification method by reducing nodes and edges based on both network and semantic structure of statistical terms.

Network structure simplification was done by limiting the number of edges between nodes. We took only top-ω edges with high co-occurrence for each nodes. For example, suppose that a statistical term S_1 co-occurred with 5 other statistical terms $\{S_{11}, S_{12}, S_{13}, S_{14}, S_{15}\}$ ordered by the co-occurrency with S_1. If we set $\omega = 3$, only edges between S_1 and $\{S_{11}, S_{12}, S_{13}\}$ stay in the network and edges between S_1 and $\{S_{14}, S_{15}\}$ are removed. This

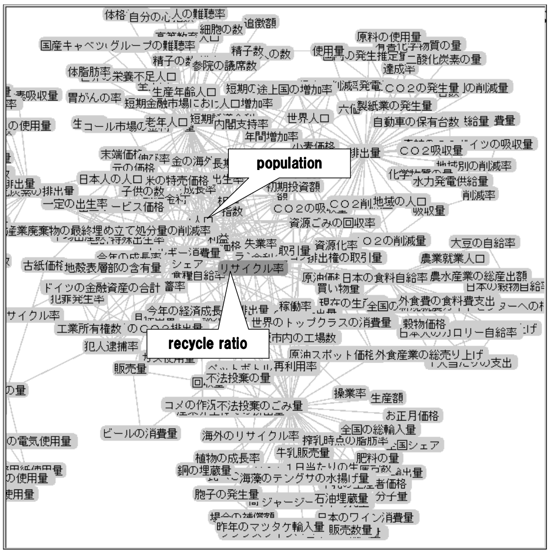

Figure 3. Statistical term network

can avoid dense clusters formed by the hubs and reveal the principal structure of the statistical term network.

Semantic structure simplification was done by clustering statistical terms considering a base form and neighbors in the network. Fig. 4 shows an overview of the proposed clustering method. In Fig. 4, two statistical terms "国内の失業率 (domestic unemployment rates)" and "アメリカの失業率 (U.S. unemployment rates)" are co-occurring with "内閣支持率 (approval rating for the cabinet)" and "経済成長率 (economic growth)." While "銃所持率 (gun ownership ratio)" is co-occurring with only "アメリカの失業率 (U.S. unemployment rates)." In this case, we can reduce edges to "内閣支持率 (approval rating for the cabinet)" and "経済成長率 (economic growth)" by clustering "国内の失業率 (domestic unemployment rates)" and "アメリカの失業率 (U.S. unemployment rates)" using their base form "失業率 (unemployment rates)" as shown in Fig. 4(b). While the relationship between "銃所持率 (gun ownership ratio)" and "アメリカの失業率 (U.S. unemployment rates)" are regionally specific topic, so it is better to keep them without modification.

Now, we can formalize our clustering method in the following steps:

1. Find two statistical terms S_1 and S_2 consisting of the same base form BF and different modifiers M, $S_1 = \{M_{11}, M_{12}, ..., M_{1m}, BF\}$, $S_2 = \{M_{21}, M_{22}, ..., M_{2n}, BF\}$.

2. If S_1 and S_2 have at least one common co-occurring term in the statistical term network, generate a new node $S_3 = S_1 \cap S_2$ and integrate edges from common co-occurring terms to S_1 and S_2.

3. If there are some nodes connecting to S_1 or S_2

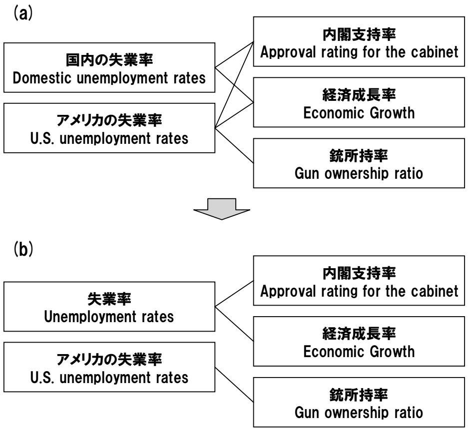

Figure 4. An overview of clustering based on network structure and base form of statistical terms

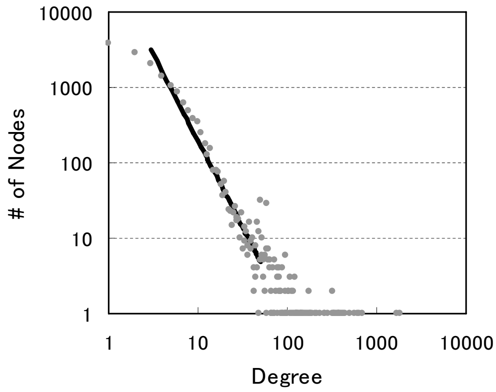

Figure 5. Degree distribution of statistical terms

independently, keep them and their edges without modification.

4 Experimental Settings

We used the MuST corpus from the NTCIR-7 Workshop provided by The National Institute of Informatics. MuST corpus consists of about 420,000 Mainichi news articles from Jan. 1998 to Dec. 2001. Some of the news articles in 27 topics, like approval rating for the cabinet and the shipping volume of PCs, are annotated texts for numeral expression extraction[4].

We used 86 statistical terms in the annotated corpus as the seed words for statistical term extraction, and expanded them one hundred times (8,600 statistical terms). For the visualization of the statistical network, we used an open source software "prefuse" provided as a visualization tool[1].

5 Results and Discussions

As a result of simplification, the relationship between statistical terms can be observed as a chain of cliques corresponding to specific topics. We will discuss the obtained statistical term network in the following sections.

5.1 Effect of Simplification

Fig. 5 shows a distribution of node degree of statistical terms, which means the distribution of the number of co-occurrence terms of a single statistical term. The solid line indicates the power law curve in the range of degree less than 50. Although there are many statistical terms with degree more than 50, most of them are too broad to show in the visualization. For example, the degree of "価格 (price)" , "株価 (share price)", "金利 (interest rates)" are 701, 607 and 511 respectively. So we removed statistical terms whose degree is more than 50 in order to visualize the relationship between more detailed statistical terms.

We also observed that higher degree limitation parameter ω can divide an original statistical term network into small clusters in which nodes are connected each other. Usually, there are the maximum cluster which includs most of statistical terms in the original statistical term network and many small clusters consists of a few statistical terms. Fig. 6 shows the maximum cluster size as a function of degree limitation parameter ω. When $\omega = 50$, the maximum cluster contained 3,862 statistical terms as nodes, and the second largest cluster was only 20 statistical terms. The rest of the nodes formed only small-sized clusters with less than 5 nodes and these clusters unconnected with each other. This tendency did not changed until $\omega = 20$, whose maximum cluster size was still 3,838. The maximum cluster shrank gradually when $10 < \omega < 20$, and it quickly turned into pieces when $\omega < 10$. Thus we chose $\omega = 10$ for the visualization because it has very simple structure and still kept 90% of the maximum cluster from $\omega = 50$.

5.2 Simplified Co-occurrence Network

Fig. 7 shows an example of the simplified co-occurrence network of statistical terms. In Fig. 7 shows related terms within two hops from the statistical term "リサイクル率 (recycle ratio)" as well as Fig. 3. It is obvious that the network density is reduced and it is much easier to observe relationships between sta-

[1] http://prefuse.org/

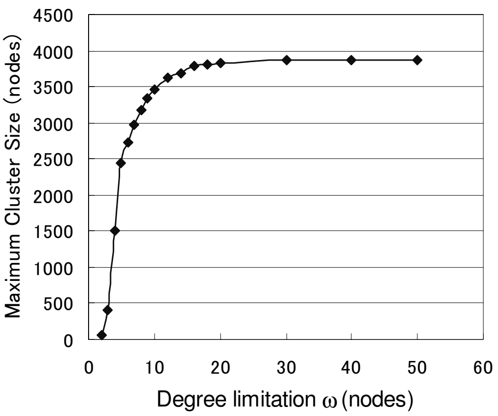

Figure 6. Maximum cluster size as a function of degree limitation ω

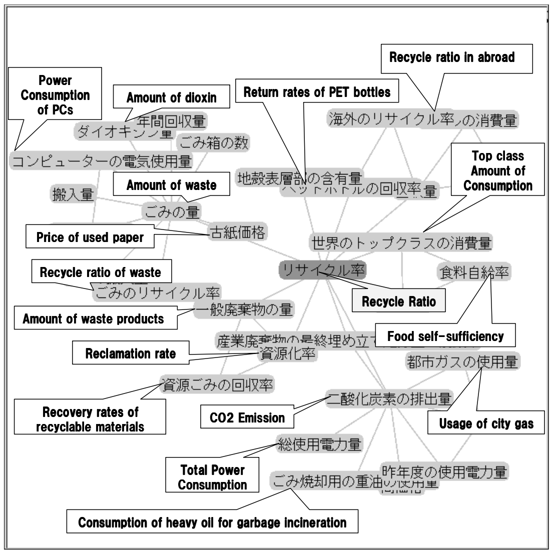

Figure 7. Simplified statistical term network (2 hops from "recycle ratio")

tistical terms. We can see the statistical term "リサイクル率 (recycle ratio)" surrounded by related statistical terms such as "ごみの量 (amount of waste)", "ペットボトルの回収率 (return rates of PET bottles)" and "二酸化炭素の排出量 (CO_2 emission)".

Fig. 8 shows the simplified statistical term network in a wide scope including related terms within 7 hops from the statistical term "リサイクル率 (recycle ratio)." In Fig. 8, we can see chained cliques connected by hubs in the network structure. We found that a clique consisted of several statistical terms extracted from the same news article, and related to a specific topic. Although these cliques do not always have causal relations, we can figure out related topics by tracking these cliques.

For example, the route (A) in Fig. 8 which runs from "リサイクル率 (recycle ratio)" to top-left shows a trend of related topics drifting to "gas emissions", "chemical emissions," "fertility," and "incidence of testicular cancer." The route (B) shows a similar topic drift to the route (A). These routes are supposed to be generated by wide attention to environmental hormone and dioxin issues in the late 90's when the news articles were written. The route (C) in Fig. 8 shows a topic drift to "propensity to consume," "consumption of rice," "food self-sufficiency," and "Farm population." The route (D) shows a spectrum from "リサイクル率 (recycle ratio)" to "unemployment rates", "economic growth", "interest rates" and "condominium price." In the route (E), after the topic slightly changed into "Resource recovery," it jumped to unexpected categories such as "sales of beer," "sales of PCs," and "market share." Although these topic drifts do not always hold logical consistency, it is possible that they may help to find unexpected relations between statistic terms and to stretch our imagination such as chance discovery[2].

5.3 Statistical Term Clustering

Table 1 shows examples of statistical term clustered by the proposed method. In many cases, common compound nouns in the suffix of statistical terms became new nodes by the clustering process such as "新規住宅着工戸数 (new privately-owned housing units started)" and "パソコン販売台数 (PC sales volume)". There are some clustered nodes which consist of both compound nouns and a possessive particle "の (of)" such as "温室効果ガスの総排出量 (total greenhouse gas emission)" and ""野菜の価格 (vegetable prices)". These results demonstrate the effectiveness of our clustering method.

There are also bad results of clustering. For example, clustering turned "ユダヤ人の数 (the number of Jewish)" into "人の数 (the number of people)", and " 事故時の状況 (situation of the accident)" and "定年時の状況 (situation at retirement)" into "時の状況 (situation at some point)". We may avoid these cases by adding a heuristic like "Do not split compound nouns for clustering."

6 Conclusions

In this paper, we proposed a simplification method of co-occurrence network structure to visualize global dynamics. We simplified the complex statistics term network by reducing nodes and edges considering both network and semantic structure of statistical terms. In the experiment conducted on news articles, we found

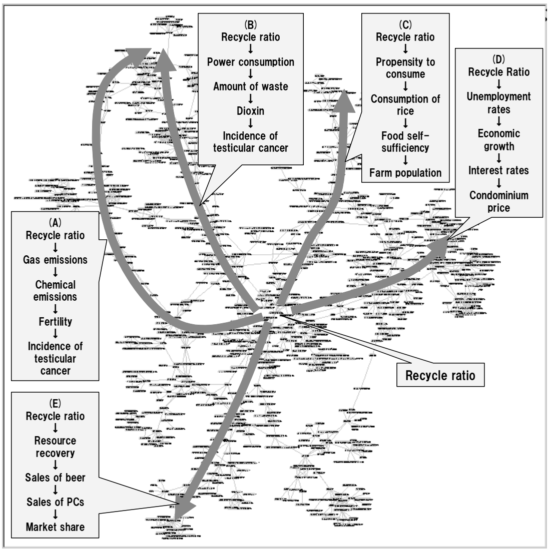

Figure 8. Simplified statistical term network (7 hops from "recycle ratio")

that chained cliques in the statistical term network can represent relationships between specific topics. In future work, we would like to develop a method to treat the direction of causalities and also make the system scalable for large corpora like the World Wide Web.

Table 1. Examples of clustered statistical terms

Statistical terms	Clustered statistical term
わが国の温室効果ガスの総排出量 Domestic total greenhouse gas emission	温室効果ガスの総排出量 Total greenhouse gas emission
96年の温室効果ガスの総排出量 Total greenhouse gas emission in 1996	
日本の温室効果ガスの総排出量 Total greenhouse gas emission in Japan	
8月の新規住宅着工戸数 New privately-owned housing units started in August	新規住宅着工戸数 New privately-owned housing units started
今年度の新規住宅着工戸数 New privately-owned housing units started this year	
12月のパソコン販売台数 PC sales volume in December	パソコン販売台数 PC sales volume
秋葉原の電気街のパソコン販売台数 PC sales volume in Akihabara	
埼玉県所沢市の野菜の価格 Vegetable prices in Tokorozawa city Saitama	野菜の価格 Vegetable prices
すべての野菜の価格 All vegetable prices	

References

[1] Katsuyuki Fujihata, Masahiro Shiga, Tatsunori Mori: Extraction of Numerical Expressions by Constraints and Default Rules of Dependency Structure, IPSJ SIG Notes, Vol.2001, No.86, pp. 119-125, 2001.

[2] Kazuhisa Inaba, Yukio Ohsawa: A Study on a Method for Supporting Scenario Extraction from Time Series Information, Proceedings of 1st Annual Workshop on Rough Sets and Chance Discovery (RSCD), 2005.

[3] Takashi Inui, Manabu Okumura: Investigating the Characteristics of Causal Relations in Japanese Text, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, 2005.

[4] Tsuneaki Kato, Mitsunori Matsushita, Noriko Kando: MuST: A Workshop on Multimodal Summarization for Trend Information, Proceedings of NTCIR-5 Workshop Meeting, 2005.

[5] Tsuneaki Kato, Mitsunori Matsushita, Noriko Kando: Expansion of Multimodal Summarization for Trend Information –Report on the First and Second Cycles of the MuST Workshop –, Proceedings of NTCIR-6 Workshop Meeting, 2007.

[6] Tatsunori Mori, Atsushi Fujioka, Ichiro Murata: Automated extraction of statistic expressions from text for information compilation, 2007-JSAI, 3H9-4, 2007.

[7] Thomas Perrin, Hideki Kawai, Kazuo Kunieda, Keiji Yamada: Global Dynamics Network Construction from the Web, Proceedings of International Workshop on Information-Explosion and Next Generation Search, 2008.

[8] Haruka Saito, Hideki Kawai, Masaaki Tsuchida, Hironori Mizuguchi, Dai Kusui: Extraction of Statistical Terms and Co-occurrence Networks from, Proceedings of NTCIR-6 Workshop Meeting, 2007.

[9] Kouich Saito, Akito Sakoda, Tomito Nakae, Yoshihiro Iwai, Naoyoshi Tamura, Hiroshi Nakagawa: Numeral Information Extraction from Newspaper's Articles, IPSJ SIG Notes, Vol. 98 No. 48, 1998.

[10] Takefumi Sato, Masahide Horita: Assessing the Plausibility of Inferense Based on Automated Construction of Causal Nnetworks Using Webmining, Sociotechnica, Vol. 4, pp.66-74, 2006.

[11] Hiroshi Sato, Kaname Kasahara, Kazumitsu Matsuzawa: Rertrieval of Simplified Causal Knowledge in Text and its Application, Technical report of IEICE. Thought and language, Vol.98, No.640, pp. 27-34, 1999.