Extraction of Statistical Terms and Co-occurrence Networks from Newspapers

Haruka Saito† Hideki Kawai‡ Masaaki Tsuchida†
Hironori Mizuguchi† Dai Kusui†
†Service Platforms Res. Lab., NEC ‡C&C Innovation Res. Lab., NEC

Abstract

In this paper, we automatically extract statistical terms and build their co-occurrence networks from newspapers. Statistical terms are expression of the measurements of statistics to watch the movements of phenomena; birth rates, public approval rating of the Cabinet and so on. In recent years, we have a vast amount of available information because of computerization and the technologies of making their overview and enhancement of their values are noticed. One of them is the technology of visualizing information of social trend and movements from newspapers. For visualizing trend information, there some approaches. In this paper, we take the approach of building networks of causal relations among the statistical terms. To extract statistical terms, we propose extraction method using suffixes. To extract causal relations among statistical terms, we first extract co-occurrence relations and next show them with the networks. We can extract many statistical terms with high accuracy by our method and find interesting links among some statistical terms by our co-occurrence networks.

1 Introduction

In a complicated modern society, we must analyze social phenomena from many angles to keep up with social trends. Many recent problems, such as environmental problems, are so complex that we cannot solve them only by searching for local solutions. In these problems, the optimum local solution is not always the best overall solution. We must integrate some information relevant to our problems and search for global optimum solutions to solve such complex problems.

On the other hand, we have a vast amount of available information because of computerization. Various social phenomena are treated as social trends in many newspaper articles. Information on social trends in newspaper articles is based on statistics, which are measurements of real world from some viewpoint. Specifically, information on social trends is often described in terms of the statistics of the trends, in terms of the events in newspaper articles, for example, “a 20% rise in birth rates,” or “the slumping public approval rating of the Cabinet because of an economic downturn.” The statistical terms are expressions of the measurements of statistics: “birth rates,” “public approval rating of the Cabinet” and so on. The terms of trends are expressions of the movements of statistics: “20% rise of” “downturn” and so on. The terms of events are expressions of phenomena relevant to the movements of statistics: “slumping,” “uprising,” “cold summer” and so on.

We can know how social phenomena exist and keep up with social trends if we can extract information on trends relevant to various social phenomena from the vast amount of newspaper articles and visualize it. There are two approaches to extracting such information and visualizing it. One is a visualization of the value of statistics based on information on trends. For example, this approach draws graphs of the temporal movements of statistics or the geographical distribution of statistics. These are approaches for extracting and visualizing information on trends focusing on one phenomenon in the real world. The other is a visualization of the relations among the statistical terms, for example extracting a causal relation among the statistical terms and showing them with networks. These are approaches for extracting and visualizing information on trends focusing on a causal relation or a generating mechanism of social phenomena in the real world.

In this paper, we take the latter approach of building networks of causal relations among the statistical terms to analyze social phenomena totally, which can allow for information on trends to be visualized. We have two main challenges in building networks of causal relations among statistical terms. One is extracting the terms of statistics automatically and the other is extracting the causal relations among the statistical terms to build their networks. To extract the terms of statistics, we adopt an extraction method using suffixes, which are typical patterns of terms of statistics. The reason why we use suffixes is that statistical terms often include particular expressions relative to the contexts of measuring statistics. In particular,

statistical terms in Japanese often have some pattern in their suffixes. To extract a causal relation among statistical terms, we first extract co-occurrence relations among the statistical terms and then extract a causal relation among the statistical terms by classifying them according to the kinds of relations. We have two reasons why we first extract co-occurence relations and classify them to extract a causal relation. One reason is that strong co-occurrence relations can have causal relations. The other reason is that there are few key expressions for annotating a causal relation. According to [6], 70% of in-text causal relations don't have explicit key expressions. So, in this paper, as a first step, we build networks of the statistical terms based on co-occurance relations, which can include causal relations. Then, we analyze how many causal relations a co-occurrence includes and how we can extract them.

2 Related Works

For extraction of the statistical terms, there are two kinds of related works. One is an extraction method of the statistical terms and the other is analysis and categorization of the statistical terms. As a method of extracting the statistical terms, Saito *et al.* [1] proposed a method using numerical expressions and their surrounding syntactic patterns, which are certain word class and particle patterns. Hujihata *et al.* [2] proposed a method selecting upper candidates that are extracted by a modification relation with numerical expressions and ordered by types of modification relation. Both [1] and [2] proposed methods using numerical expressions because their aims are the extraction of pairs of statistical terms and numerical expressions. On the other hand, our method uses suffixes attached to statistical terms because our aim is extraction of a causal relation among the terms of statistics. The reason why we adopt the approach is that terms of statistics do not always appear with numerical expressions, and we want to extract as many statistical terms as possible. For example, in "the rate of unemployment is increasing," there are terms of statistics without numerical expressions, and we don't want to miss such terms. For analysis and categorization of statistical terms, Murata *et al.* [3] classified statistical terms to make a training data set for automated extraction of statistical terms. On the other hand, we classify statistical terms to extract a causal relation among the statistical terms considering their hierarchy.

For extraction of a causal relation, [4][5] proposed a method for extracting an explicit causal relation by conjunction expressions or case frame dictionaries, such as "because" and "since". However, Inui *et al.* [6] reported results of examinations that over 70% of a causal relation doesn't have such explicit expressions of a causal relation. So, we try to extract a causal relation by means of extracting co-occurrence relations and classifying them.

3 Extraction and Classification of statistical terms

In this section, we propose a method of extracting statistical terms using suffixes, which are typical patterns of terms of statistics. Then, we give labels of classified and categorized statistical terms to build a causal relation considering hierarchy.

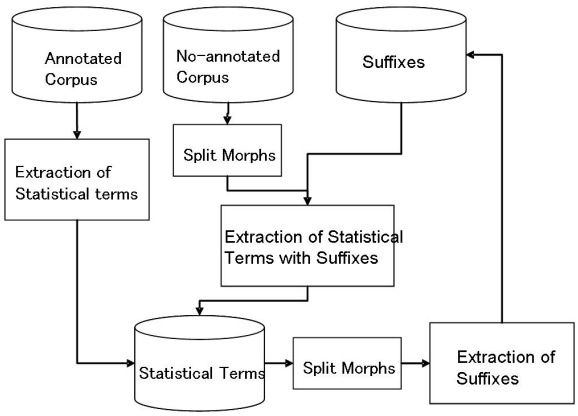

Figure 1. The procedure for extracting statistical terms

3.1 Extraction of statistical terms with Suffixes

We extract many statistical terms from corpuses using a few suffixes. To be more precise, we first have some seed statistical terms, and we extract noun phrases whose bottoms are suffixes as increased terms of statistics from corpuses. These suffixes are typical patterns of seed terms of statistics. Fig. 1 shows our procedure for extracting terms of statistics. First, we extract tagged words as seed statistical terms from annotated corpuses and store them in a dictionary of terms of statistics. Next, we split seed statistical terms into morphs and extract from one to three bottom morphs as suffixes. Finally, we extract noun phrases whose bottoms are suffixes as increased terms of statistics from non-annotated corpuses and store them in the dictionary.

The seed statistical terms are strings between "<unit stat>" tags or "<name>" tags. According to the specification of MuST tags [7], "<unit stat>" tags express parts of referred to statistics to be visualized and "<name>" tags also express names of statistics. These correspond to the statistical terms we defined.

So in this paper, we regard strings between the tags as terms of statistics. In the step of extraction with

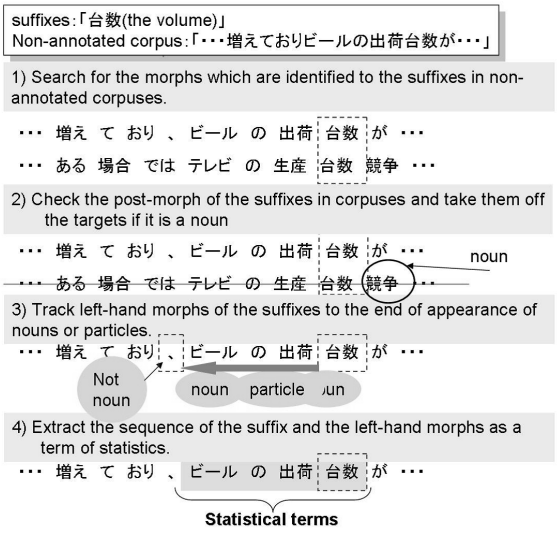

Figure 2. The algorithm of extraction of statistical terms

suffixes, we extract noun phrases that include their modifier and whose bottoms include the suffixes as increased statistical terms. To be more precise, we search for the morphs identified in the suffixes and we extract them and their left-hand sequence of nouns or particles as statistical terms. We explain the details of the algorithm in Fig. 2. Fig. 2 shows an example whose suffix is "台数 (volume)" and whose parts of sentences in a non-annotated corpus are "・・・増えており、ビールの出荷台数が・・・(・・・ are increasing, and the volume of beer shipments is ・・・)" and "・・・この場合はテレビの販売台数競争・・・(・・・ in this case, the competition in sales for TVs ・・・)." The process is as follows.

1. Search for the morphs that are identified with the suffixes in non-annotated corpuses. In Fig. 2, both sentences include the suffix "台数 (volume)."

2. Check the post-morphs of the suffixes in corpuses and take them off the targets if they are nouns. In the latter sentence of Fig. 2, there is a noun "競争 (competition)" at the post-suffix.

3. Track left-hand morphs of the suffixes to the end of the appearance of nouns or particles. In Fig. 2, the left-hand morphs of the suffix "台数 (volume)" are "出荷 (shipments)) [noun]", "の (of)[particle]", "ビール (beer)[noun]", "、(,)[code]", $cdots$ in sequence. The tracking of morphs is over when "、(,)" appear. We want to extract sequences of nouns and particles because those must be noun phrases.

4. Extract the sequence of the suffix and the left-hand morphs as a term of statistics. In Fig. 2, tracking is over when the left-hand morph of the suffix is "、(,)". So we extract the sequence morphs from the next morphs of "、(,)" to the suffix "台数 (volume)", "ビールの出荷台数 (volume of beer shipments)," as statistical terms.

However, we don't extract the statistical terms at 4 whose first morph is "の (of)" because they are not noun phrases.

We also define a score of statistical terms that is based on numbers of morphs matched to suffixes and their occurance frequencies. We expect the score to be viable for selecting proper statistical terms. Characteristics of the proper statistical terms are (1) their suffixes have many morphs, and (2) their suffixes have high occurrence frequencies. That is, the score S of statistical terms whose suffix has N morphs and whose occurrence are R times is:

$$S = 100^{N-1} \times R.$$

3.2 Classifying statistical terms

We classify the extracted statistical terms by regarding their modifiers to analyze their abstraction levels. While [3] also investigated classifying statistical terms, we classify terms of statistics to build causal networks among them. All statistical terms with consist of one causal network should heve the same abstraction level. For example, there are some statistical terms, "失業率 (rate of unemployment)", "アメリカの失業率 (American rate of unemployment)", "国内の失業率 (domestic rate of unemployment)" and so on, whose suffix is "率 (rate)". In this case, "失業率 (rate of employment)" and "アメリカの失業率" each need different statistical terms to build their causal networks. Building a causal network including "失業率 (rate of employment)" requires statistical terms whose abstraction levels are high, while building a causal network including "アメリカの失業率 (American rate of employment)" requires statistical terms whose abstraction levels are not so high and include the attribute American.

For this purpose, we define four labels of modifiers to classify terms of statistics based on the following viewpoints: (1) modifiers belonging to the four labels should appear in many statistical terms mutually; (2) the multiple modifiers belonging to the same labels never appear in one statistical term. Modifiers that do not belong to the four labels are classified as "others." Fig. 3 illustrates the four labels, explained as follows.

- **fundamental statistical terms** Noun phrases of statistical terms whose number of morphs are

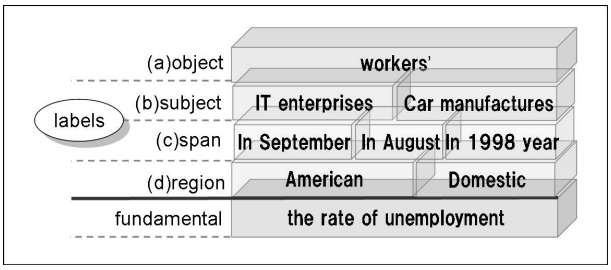

Figure 3. Labels of statistical terms

minimum and whose context is detectable. In Fig. 3, "失業率 (rate of unemployment)" is a fundamental statistical term.

(a)object Modifiers including statistical terms that express target people, organizations, and things to be measured in statistics. In Fig. 3, "workers"' is a modifier belong to the object label.

(b)subject Modifiers including statistical terms that express people, organization, and things that can operate on or control the statistics. In Fig. 3, "IT enterprises"' and "auto manufactures"' are modifiers belonging to the subject label.

(c)span Modifiers including statistical terms that express spans of time measured in the statistics. In Fig. 3, "in September", "in August", and "in 1998" are modifiers belonging to the span label.

(d)region Modifiers including statistical terms that express regions or areas measured in the statistics. In Fig. 3, "American" and "domestic" are modifiers belonging to the region label.

We now explain the difference between **(a)object** and **(b)subject**. The former is things that are measured as objects of statistics, and the latter is things that can change statistics. For example, "workers" can be measured as objects of the rate of unemployment but cannot control the rate of unemployment. While "IT enterprises" can make decisions about employment and can make an impact on the rate of unemployment.

The definition of these labels can reveal the abstraction levels of statistical terms. That is, we find that the more kinds of labels a statistical term has, the lower the abstraction level and the more complex the statistical term is. For example, compared with "the rate of workers' unemployment", "the domestic rate of workers' unemployment in September" is a lower abstraction level and a more complex statistical term.

4 Networks of Co-occurrent Statistical Terms

In this section, we explain the method for building networks of co-occurrence statistical terms.

In this paper, we interactively make networks to manually select necessary relations of statistical terms that are subsets of system output co-occurrent statistical terms. The reason why we adopt an interactive method for building is that we want to extract only terms of statistics that have a causal relation from statistical terms that have co-occurrence relation.

We define two statistical terms as having a co-occurrence relation or as being co-occurrent when they appear in one paragraph together. We also define the co-occurrence frequency of two statistical terms as the number of paragraphs in which they appear together.

As follows, we explain our method for interactively building networks. At first, we extract the co-occurrent statistical terms. In particular, we label each statistical term with a paragraph ID of its appearance, and then we extract pairs of statistical terms that have the same ID as co-occurrent statistical terms.

Next, we remove statistical terms that are co-occurrent with many statistical terms. Otherwise, when we build networks whose nodes are statistical terms, these nodes become hubs that connect most nodes. In this paper, statistical terms that are co-occurrent with over 50 statistical terms are removed. For example, "fraction", "recurring profits" are removed because they are each co-occurrent with over 50 statistical terms. Finally, we interactively expand the networks. We iteratively display one hop co-occurrent statistical terms with our focused statistical term, select the interesting statistical terms, and again display their one hop co-occurrent statistical terms. We exit this interactive process when most outside nodes do not have added co-occurrent statistical terms.

5 Results and Discussion

In this section, we show the extraction method using suffixes enabling the extraction of many statistical terms. We also show that most were statistical terms labeled "span," the second most were labeled "object," and a few statistical terms had multiple labels. In building networks of co-occurrent statistical terms, we show we could find some relations of statistical terms that are difficult to suggest. We also show there were many pairs of co-occurrent statistical terms whose relations were direct and indirect relations when we classified statistical terms that had co-occurrence relation.

Table 1. Statistical terms in annotated articles

statistical terms in annotated articles
完全失業者数 (the rate of unemployment)
PHS の加入台数 (the volume of PHS enrollment)
国内出荷台数 (the domestic volume of shipments)
実質消費支出 (real consumption expenditure)
花粉の飛散量 (the amount of pollen dispersal)

Table 2. Examples of suffixes and their frequencies

suffixes of statistical terms	frequency
数 (number)	12
率 (rate)	6
相場 (market)	4
指数 (index number)	4

5.1 Results of Extraction of Statistical Terms

5.1.1 Methods of Experiments

We experimented with our method of extracting statistical terms using suffixes from a newspaper corpus. We used newspaper corpuses of 1998 and 1999 from daily newspapers including about 23,000 articles. The 23,000 articles consist of about 1,000 annotated articles and about 22,000 non-annotated articles. The 86 terms were extracted from the annotated articles as seed statistical terms.

In Table 1, we show some examples of seed terms of statistics extracted from the annotated articles. In Table 2, we show some examples of 146 suffixes that were extracted from the seed statistical terms and sorted by occurrence frequency. We extracted and kept the scores of statistical terms by the extraction method of section 3.1 using these suffixes.

We manually evaluated the accuracy of 200 candidate statistical terms which were chosen randomly from top 16000 score ranking and 1000 candidates which were top 1000 score ranking. The former were candidate words chosen randomly from 1600 candidate statistical terms whose scores were over 5. The later were candidate words whose scores were in the top 1,000 ranking. In this evaluation, three people manually check whether each candidate is a real statistical term or not. Candidate words that had different judgments by the judges were categorized as "not statistical terms" and only candidates all judges deemed "real statistical terms" were categorized as correct terms of statistics. The accuracy was stimulated by the rate of words that were correct statistical terms.

Table 3. Statistical terms extracted using suffixes (top 5 scores)

statistical terms	score
完全失業率 (the rate of unemployment)	3,620,000
日経平均株価 (the average of share price)	3,310,000
完全失業者数 (the rate of unemployment)	1,560,000
内閣支持率 (cabinet support rate)	1,000,000
有効求人倍率 (active opening ratio)	920,000

5.1.2 Results

We could extract 33,100 candidate statistical terms by our proposed method and the accuracy of 200 candidates chosen randomly from 16,000 was 0.67. The accuracy of 1,000 candidates in the top ranking was 0.84. The percentage of the coincidence of judgments was 87% for about 200 candidates chosen randomly and 88% for candidates ranked in the top 1,000. Therefore, we conclude that our extraction method is effective for increasing statistical terms.

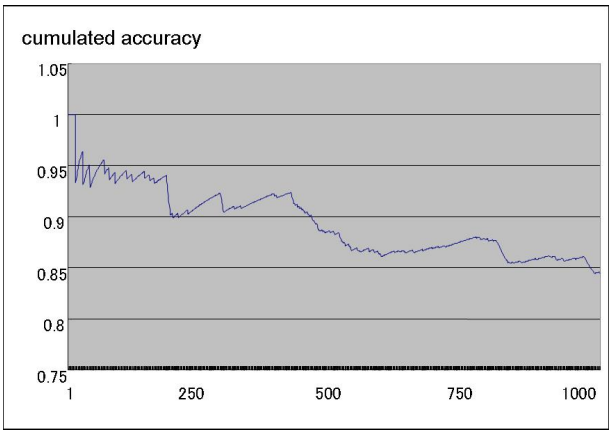

Figure 4. Accuracy results of top 1000

Table 3 shows examples of statistical terms whose scores are in the top 10 ranking. Fig. 4 shows the result of a cumulated accuracy graph of top 1000 ranking statistical temrs. In Fig. 4, the x-axis indicates the score ranking and the y-axis indicates cumulated accuracy from top to the x score ranking. Fig. 4 shows that the cumulated accuracy goes down as the score ranking does. The rate of accuracy was 0.84 for those whose score ranking was in the top 1,000.

The error examples were "最悪の失業率 (the worst rate of unemployment)", "一回の割合 (the rate of once)", "年間最高の 127 万の患者 (the largest number 1.27 million of patients for a year)", "学校数や在学者数 (the number of schools and students)", and so on.

Table 4. Classification of statistical terms (top 200)

label(number)[%]	examples
(a)object(48)[24]	PHS の加入台数 (the volume of PHS enrollment), ダイヤの売上高 (the sales of diamond), 橋本内閣支持率 (Hasimoto cabinet support rate)
(b)subject(4)[2]	百貨店売上高 (the sales of department), 高校の中退者数 (the number of students dropped out of high school)
(c)span(97)[49]	二月の国内卸売り物価指数 (domestic price index numbers of wholesaling in February)
(d)region(38)[19]	二月の国内卸売り物価指数 (domestic price index numbers of wholesaling in February)
fundamental(18)[9]	支持率 (support rate), 観客数 (the size of audience)
others(18)[9]	地域別の完全失業率 (the rate of unemployment by area)

5.1.3 Discussions

In this experiment, we use suffixes to extract statistical terms. Although their accuracy is high, the issue of completeness remains. There are some statistical terms not included in suffixes, such as "GDP." We have to study other extraction methods for completeness.

In this experiment, targets of extraction were limited to terms of statistics. However, terms of events, "uprising", "cold summer" and so on, are relative to movements of social phenomena other than terms of statistics. We need terms of events because they often have a causal relation with statistical terms. It is difficult to apply our extraction methods using suffixes for extraction of terms of events because it is difficult to find feature strings with alternative suffixes. We have to study other extraction methods for extracting information other than statistical terms.

5.2 Experiments of Classifying

5.2.1 Method of Experiments

We manually classified the 200 statistical terms that were extracted by our method and were correct. Our labels defined in section 3.2 were used for the classification.

5.2.2 Results

Table 4 shows the result of classifying 200 terms of statistics. The values inside parentheses indicate the pertinent number of statistical terms. In Table 4, we found that the statistical terms including "span"-labeled modifiers were the most frequent and "region"-labeled the least. In the 180 statistical terms that had some labels, 156 (87%) terms of statistics had only one kind of label each, 23 (13%) terms of statistics had two kinds of labels each, and 1 (0.5%) term of statistics had three kinds of labels. These results showed that the top 200 ranking statistical terms included few complex statistical terms whose modifiers belonged to several labels.

5.2.3 Discussion

The result showing few complex statistical terms demonstrates that the extraction method using suffixes cannot apply to extracting complex terms of statistics whose modifiers belong to several labels. Complex statistical terms have many modifiers but they rarely include all modifiers in one noun phrase. Their many modifiers are often distributed in titles or previous sentences. So we need to propose another method that can extract complex statistical terms by collecting co-occurrence statistical terms that whose fundamental statistical terms are common and appearence positions are close and extracting their modifiers.

5.3 Results of Building Networks

5.3.1 Methods of Experiments

We built networks of co-occurrent statistical terms by the interactive method introduced in section 4. To build them, we used 33,100 statistical terms extracted in section 5.1 and 22,000 non-annotated articles. The 33,100 terms included incorrect statistical terms. The graphic tool in this experiment was a tool originally developed at NEC Labs [8].

5.3.2 Results

Fig. 5 is the result of a co-occurrence network starting from the term "birth rates," which has 3 hops. Each node is a statistical term and each pair of nodes connected in an arc is a co-occurrence relation.

5.3.3 Discussion

In Fig. 5, there were interesting statistical terms, "the rate of recycling" and "emissions of carbon dioxide" as terms linked to "the rate of birth." In studying the relations between the rate of birth and each part of the network by reading the article, the left part included "the number of museums and bookstores", "the number of facilities of overtime childcare", and "the number of emergency hospitals", and had co-occurrence but did not have a causal relation with "the rate of birth". These statistical terms which were the left part

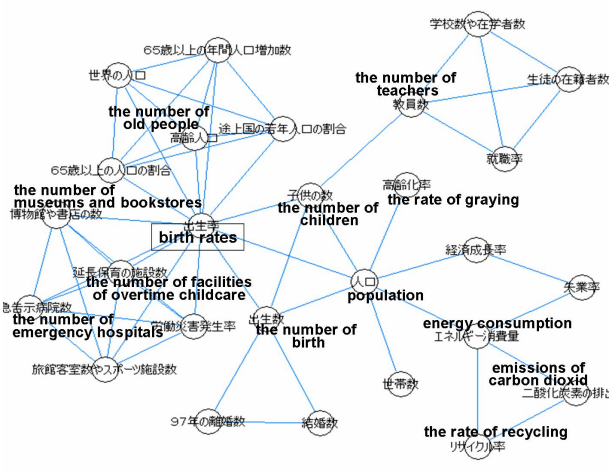

Figure 5. Co-occurrence network starting from "the rate of birth"

of the network were indicators of richness of prefectures. The right part including "population", "energy consumption" and so on had a causal relation with "the rate of birth." We expect to extract interesting relations among statistical terms when we focus attention on a causal relation from these co-occurrence relations. We need works to extract a causal relation from co-occurrence relations in the future.

In this paper, networks were built of statistical terms as nodes, but interesting relations among phenomena were often connected by statistical terms. So we also need works to build networks considering terms of events.

5.4 Results of Classifying Co-occurrence Relations

5.4.1 Methods of Experiments

We classified 200 statistical terms that were in the top 200 ranking of co-occurrence tendencies. The classification was conducted manually by three judges. The co-occurrence tendencies are the rate of the number of co-occurrence appearences of the pair in the less number of appearencies of the pair. We prepared four kinds of labels: "synonym relations", "direct relations", "indirect relations" and "no relations." We define a direct relation between two terms as having a dependence relation between them and an indirect relation between two terms as having an independence relation but their common term has a dependence relation with both of them. The judges decided based on their own knowledge what kind of labels 200 pairs of the statistical terms belonged to each. They also checked by reading the 614 articles where the pairs appeared to study difficult levels for deciding kinds of ralations of the pairs. The difficult level 1 was the case where the pairs had the explicit clue expressions about their relations in the articles. The difficult level 2 was the case where the pairs had not the explicit clue expressions about their relations but the clue articles for deciding their relations. The difficult level 3 was the case where the pairs had neither explicit clue expressions about thier relations nor the clue articles of their relations. We adopted majority judgments when judgments were different.

5.4.2 Results

The 200 pairs that had high co-occurrent tendencies included direct and indirect relations in this order and half of them were described in their relations in the articles. However, there were few clue expressions. The judges' judgments coincided 48% of the time if the relations were described in the articles and 49% of the time if clue expressions were described in the articles. Table 5 shows details of the results of the relation classifications. It indicates the number of pairs of each relation, pairs of difficult level 1, level 2, and level 3. The "unknown" relation is a judgment for which each judge had a different result. The numbers inside parentheses indicate the rate. That is, the first column indicates the rate in 200 pairs. The second column indicates the rate of each relation. For example, there were 119 pairs that had a direct relation in all 200 pairs, so the rate was 60%. In the 119 pairs, there were 68 pairs describing their relations in the articles, so the rate was 57%.

This table shows that there were many pairs that had a direct relation when their co-occurrence tendencies were high, and half of them were described their relations in the articles but few of them had clue expressions, that is, the most of them were the difficult level 2. On the other hand, most synonym relations were described in the articles and had clue expressions, that is, the most of them were the difficult level 1.

Examples of clue expressions were "because" or "result of" for direct relations; "behind" for indirect relations; and "( )*parenthesis*" or "indicate" for synonym relations.

We also analyzed cases where relations were described in the articles or not. We found that easy relations, "the rate of unemployment and the number of people unemployed" and so on, were not described. On the other hand, we found that difficult relations, such as "large electric consumption and sales of departments", were described explicitly. The reason seems to be that the authors of the articles chose the explicit description of relations considering the level of the readers.

5.4.3 Discussions

The results show there were 173 (87%) pairs that had direct or indirect relations (some relations) in co-

Table 5. Results of classifying co-occurrence relations

labels	number[%]	level 1	level 2	level 3
direct	119[60]	32	36	51
indirect	54[27]	7	16	31
synonym	12[6]	10	0	2
no	12[6]	-	-	-
unknown	3[1.5]	-	-	-

occurrent relations. However, there were 39 pairs that had clue expressions in them and we guess there will be few clue expressions in causal relations appearing in the articles, so we guess there were few clue expressions, such as conjunction expressions, for example. On the other hand, for synonym relations, there were many clue expressions, for example, parentheses, which were included in the description of relations in the articles. Therefore, we can extract a causal relation with an accuracy of 92% by removing terms whose relations were synonym terms from high co-occurrence incidents using clue expressions.

6 Conclusions

In this paper, we proposed that a causal relation among terms of statistics should be displayed with networks to discern movements of social phenomena from various viewpoints. At first, we extracted 1000 statistical terms with an 84% accuracy. We also classified them to analyze their abstraction and found that there were few statistical terms for which abstraction levels were low. Next, we displayed with networks the co-occurrence relations among the terms of statistics that were extracted, and analyzed cases in which we could find interesting relations among statistical terms or not. We also classified co-occurrence relations and showed they included many direct and indirect relations but few clue expressions. We will extract causal relations between terms of events and terms of statistics by notations of conjunction expressions and movement expressions in the future.

References

[1] Kouich Saito, Akito Sakoda, Tomito Nakae, Yoshihiro Iwai, Naoyoshi Tamura, Hiroshi Nakagawa. “Numeral Information Extraction from Newspaper’s Articles.” Vol. 1998 No. 48, 1998-NL-125, (1998).

[2] Katsuyuki Hujihata, Masahiro Shiga, Tatsunori Mori. “Extraction of Numerical Expressions by Constraints and Default Rules of Dependency Structure.” Vol. 2001 No. 86 2001-NL-145, (2001).

[3] Ichiro Murata, Tatsunori Mori. “Automated Extraction of Statistical Name with Machine Learning from Newspapers.” NTCIR-5 Pilot Workshop MuST, (2006).

[4] Hiroshi Satoh, Kaname Kasahara. “Acquisition of surface causal knowledge in text and Their Application.” TL-98-23, pp.27–34, (1998).

[5] Takeshi Satoh, Masahide Hotta. “Automated Building Causal Relations Networks with Web Mining.” Social Technical papers, Vol. 4, pp. 66–74, (2006).

[6] Takashi Inui, Manabu Okumura. “Characteristics of In-text Causal Relations.” IPSJ Vol. 2005 No. 50, (2005).

[7] Mitsunori Matsushita, Tsuneaki Kato. “Basic Discussions of Information Visualization Based on Movement Information.” 2005-JSAI, 1E3-03, (2005).

[8] Hironori Mizuguchi, Dai Kusui, Taku Ohoshima, Shigehiko Kanaya, Hirotada Mori. ”KAREI-DMAP: A System for Predicting and Mining Gene Regulatory Networks.” Genome Informatics 14: 382-383, (2003).