PAPER Special Section on Knowledge-Based Software Engineering

Cause Information Extraction from Financial Articles Concerning Business Performance

Hiroyuki SAKAI[†a)], *Nonmember and* Shigeru MASUYAMA[†b)], *Member*

SUMMARY We propose a method of extracting cause information from Japanese financial articles concerning business performance. Our method acquires cause information, e.g. "自動車の売上げが好調 (*zidousya no uriage ga koutyou*: Sales of cars were good)". Cause information is useful for investors in selecting companies to invest. Our method extracts cause information as a form of causal expression by using statistical information and initial clue expressions automatically. Our method can extract causal expressions without predetermined patterns or complex rules given by hand, and is expected to be applied to other tasks for acquiring phrases that have a particular meaning not limited to cause information. We compared our method with our previous one originally proposed for extracting phrases concerning traffic accident causes and experimental results showed that our new method outperforms our previous one.

key words: *cause information, causal expression, knowledge extraction, information extraction*

1. Introduction

We propose a method of extracting cause information from Japanese financial articles concerning business performance. Our method extracts phrases implying cause information, e.g. "自動車の売上げが好調 (*zidousya no uriage ga koutyou*: Sales of cars were good)" or "鉄管の売上げが不振 (*tekkan no uriage ga husin*: Sales of iron tubes were down)". Here, we define a phrase implying cause information as a "causal expression".

Collecting information concerning business performance is a very important task for investment. If the business performance of a company is good, the stock price of the company will rise in general. Moreover, cause information of the business performance is also important, because, even if the business performance of a company is good, its stock price will not rise if the main cause is the recording of an extraordinary profit not related to core business (e.g. profit from sales of stocks). This is also the case for the bad business performance. Hence, cause information of the business performance is useful for investors in selecting companies to invest. However, since there are a number of companies that announce business performance, acquiring their all cause information manually is a considerably hard task. Hence, we propose a method of identifying articles concerning business performance and extracting cause information from them automatically by using statistical information.

We consider that various kinds of important information for the investment exist besides cause information of the business performance. For example, information concerning business tie-up or company bankruptcy is significant for investment. In this paper, we adopt cause information of the business performance as information to be extracted for the following reason: Since our method extracts causal expression by using statistical information without predetermined patterns or complex rules given by hand, a number of documents concerning business performance are necessary. Moreover, it is easy to collect articles concerning business performance since they are described to each company every year. In the future, we expect that the automatic trading system that identifies articles concerning business performance and whether cause information has an influence on stock prices is able to be constructed by utilizing causal expressions extracted by our method. Hence, we adopt cause information of the business performance as information to be extracted.

First, our method extracts articles concerning business performance from newspaper corpus as a preparation. Next, our method extracts causal expressions automatically from these articles by using statistical information and two initial clue expressions. Here, a clue expression is defined as a phrase frequently modified by causal expressions. For example, "が好調 (*ga koutyou*: is good)" is a clue expression since it is frequently modified by causal expressions. Our method extracts an expression that consists of a clue expression and a phrase that modifies it as a causal expression. Hence, if many clue expressions effective for extracting causal expressions are acquirable, causal expressions are able to be extracted automatically. However, it is hard to acquire many clue expressions effective for extracting causal expressions by hand. Hence, our method also acquires such clue expressions automatically from a set of articles concerning business performance.

We introduce our method in Sects. 2 and 3. Here, the method for extracting articles concerning business performance is introduced in Sect. 2 and the method for extracting causal expressions is introduced in Step 3, respectively. Experimental results of evaluation are reported in Sect. 4 and analysis of the results on the experiments are discussed in Sect. 5. We describe related work in Sect. 6 and explain difference between our proposed method and the related work. Sect. 7 concludes this paper.

Manuscript received July 2, 2007.
Manuscript revised October 17, 2007.

[†]The authors are with the Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
a) E-mail: sakai@smlab.tutkie.tut.ac.jp
b) E-mail: masuyama@tutkie.tut.ac.jp

DOI: 10.1093/ietisy/e91–d.4.959

Table 1 Examples of selected features.

売り上げ (*uriage*: sales)	見通し (*mitooshi*: forecast)	経常利益 (*keizyou rieki*: current profits)
好調 (*koutyou*: good)	黒字 (*kurozi*: surplus)	人件費 (*jinkenhi*: employment cost)
利益 (*rieki*: profit)	純利益 (*zyunrieki*: absolute profit)	特別損失 (*tokubetu sonshitu*: extraordinary loss)
不振 (*hushin* down)	赤字 (*akazi*: deficit)	上方修正 (*zyouhou syuusei*: upward adjustment)

2. Extraction of Articles Concerning Business Performance (Preprocessing)

As a preprocessing, our method extracts articles concerning business performance from newspaper corpus by using Support Vector Machine (SVM) [1]. As training data, we manually extract 2,920 articles concerning business performance as positive examples and 2,920 articles not concerning business performance as negative examples from Nikkei newspapers published in 2000. Here, some of words contained in the positive examples are used as features of SVM. The method for extracting content words effective as features is as follows:

First, our method calculates score $W(t_i, S_p)$ of word t_i contained in positive example set S_p and score $W(t_i, S_n)$ of word t_i contained in negative example set S_n by the following Formula 1.

$$W(t_i, S_p) = P(t_i, S_p)H(t_i, S_p), \tag{1}$$

where, $P(t_i, S_p)$ is the probability that word t_i appears in positive example set S_p and is calculated by the following Formula 2

$$P(t_i, S_p) = \frac{Tf(t_i, S_p)}{\sum_{t \in Ts(S_p)} Tf(t, S_p)}, \tag{2}$$

Here, $Tf(t_i, S_p)$ is the frequency of word t_i in positive example set S_p and $Ts(S_p)$ is the set of words contained in S_p.

$H(t_i, S_p)$ is the entropy based on the probability $P(t_i, d)$ that word t_i appears in document $d \in S_p$. The entropy $H(t_i, S_p)$ is calculated by the following Formula 3:

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d), \tag{3}$$

$$P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d' \in S_p} tf(t_i, d')}, \tag{4}$$

where, $tf(t_i, d)$ is the frequency of word t_i in document d.

Next, our method compares $W(t_i, S_p)$ with $W(t_i, S_n)$. If score $W(t_i, S_p)$ is larger than $2W(t_i, S_n)$, word t_i is selected as a feature of SVM. Entropy $H(t_i, S_p)$ is introduced for assigning a large score to a word that appears uniformly in each document contained in positive example set S_p. For example, when word t_i is contained only in one document, $H(t_i, S_p) = 0$. Although such a word t_i may be an important word for the document, it may be an irrelevant word for positive example set S_p. Hence, word t_i with small entropy value should not be selected as a feature. In contrast, a word that appears uniformly in documents contained in document set S_p has a large entropy value. However, Formula 1 may assign a large score to a general word not relevant to business performance. Such a general word may also be assigned a large score in the negative example set. Hence, in our method, not only $W(t_i, S_p)$, a score in positive example set S_p, but also $W(t_i, S_n)$, a score in negative example set S_n, are calculated and compared. Some examples of selected features are shown in Table 1.

3. Extraction of Causal Expressions

Our method extracts causal expressions from articles concerning business performance extracted by the method described in Sect. 2. Here, a causal expression is a part of a sentence consisting of some "*bunsetu*'s" (a *bunsetu* is a basic block in Japanese composed of several words). Our method extracts causal expressions by using clue expressions, i.e. phrases frequently modified by causal expressions. For example, a causal expression concerning good business frequently modifies clue expression "が好調 (*ga koutyou*: is good)" and a causal expression concerning bad business frequently modifies clue expression "が不振 ([*ga husin*: is down)" in Japanese. Our method extracts an expression that consists of a clue expression and a phrase that modifies it as a causal expression. Hence, if many clue expressions effective for extracting causal expressions is acquirable, causal expressions are extracted automatically. However, it is hard to acquire many clue expressions effective for extracting causal expressions by hand. Hence, our method also acquires such clue expressions automatically from a set of articles concerning business performance.

3.1 Overview of Causal Expressions Extraction

Our method for acquiring clue expressions is as follows.

Step 1: Input a few initial clue expressions and acquire phrases that modify them. Here, we use two clue expressions, "が好調 (*ga koutyou*: be good)" and "が不振 (*ga husin*: be down)", as initial clue expressions.

Step 2: Extract phrases appearing frequently in a set of the phrases acquired in Step 1 (e.g. 売り上げ (*uriage*: sales)). In this paper, such a phrase extracted in Step 2 is defined as a "frequent expression".

Step 3: Acquire new clue expressions modified by the frequent expressions.

Step 4: Extract new frequent expressions from a set of phrases that modify the new clue expressions acquired in Step 3. This step is the same as Step 2.

Step 5: Repeat Steps 3 and 4 until they are executed predetermined times or neither new clue expressions nor new frequent expressions are extracted.

An outline of the method is shown in Fig. 1. Moreover, our method eliminates inappropriate clue expressions by using statistical information in the set of articles concerning business performance and the set of articles not concerning business performance.

3.2 Extraction of Frequent Expressions

A frequent expression is defined as a phrase appearing frequently in a set of the phrases that modify clue expressions. For example, in the following causal expressions, "A の売り上げが好調 (*A no uriage ga koutyou*: sales of A were good)", "B の売り上げが不振 (*B no uriage ga husin*: sales of B were down)", "C の売り上げが落ち込んだ (*C no uriage ga otikonda*: sales of C were down)", the frequent expression is "売り上げ (*uriage*: sales)" in these examples. Although the frequent expression consists of a word in this case, a frequent expression may consist of several words (e.g., 利益率の高い自社製品 (*riekiritu no takai zisya seihin*: products with high profit rate). The method for extracting "frequent expressions" from a set of phrases that modify clue expressions is as follows.

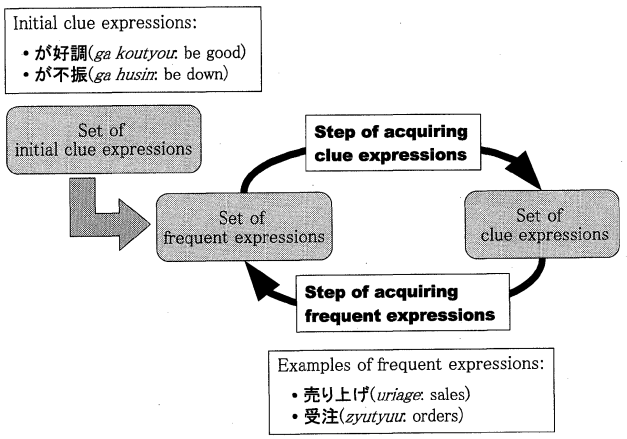

Fig. 1 Outline of our method.

Step 1: Acquire a *bunsetu* modifying a clue expression and eliminate a case particle from the *bunsetu*. We define the resulting noun as c.

Step 2: Acquire frequent expression candidates by adding *bunsetu* modifying c to c. (See Fig. 2.)

Step 3: Calculate score $S_f(e, c)$ of frequent expression candidate e containing c by the following Formula 5.

Step 4: Adopt e assigned the best score $S_f(e, c)$ among the set of frequent expression candidates containing c as a frequent expression.

Score $S_f(e, c)$ is calculated by the following Formula 5:

$$S_f(e, c) = -f_e(e, c)\sqrt{f_p(e)}\log_2 P(e, c), \tag{5}$$

where,

$P(e, c)$: the probability that frequent expression candidate e containing c appears in the set of articles concerning business performance.

$f_e(e, c)$: the number of frequent expression candidate e's containing c in the set of articles concerning business performance.

$f_p(e)$: the number of *bunsetu*'s that compose e.

$P(e, c)$ is calculated by the following Formula 6.

$$P(e, c) = \frac{f_e(e, c)}{Ne(c)}, \tag{6}$$

where, $Ne(c)$ is the total number of frequent expression candidates containing c in the set of articles concerning business performance. Examples of score $S_f(e, c)$ assigned to frequent expression candidates are shown in Table 2. In this case, "売り上げ (*uriage*: sales)" is adopted as a frequent expression.

3.3 Selection of Frequent Expressions

The frequent expressions extracted from a set of phrases that modify clue expressions may contain inappropriate ones. Hence, our method selects appropriate frequent expressions from them. Here, our method calculates entropy $H(e)$ based on the probability $P(e, s)$ that frequent expression e modifies clue expression s and selects frequent expressions assigned

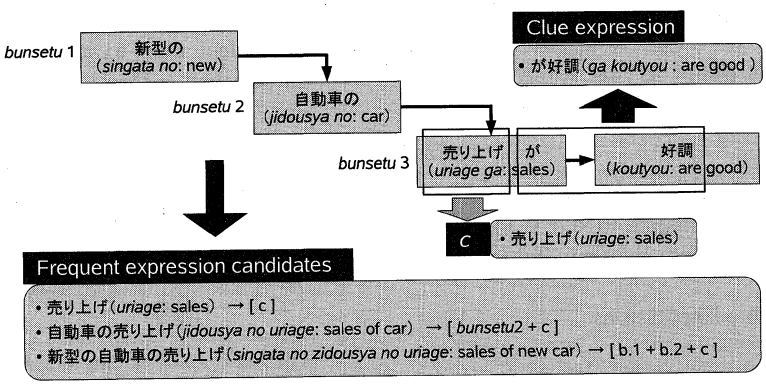

Fig. 2 Examples of frequent expression candidates.

Table 2 Examples of score assigned to frequent expression candidates.

frequent expression candidate	$f_e(e,c)$	$S_f(e,c)$
売り上げ (*uriage*: sales)	634	1063.1
半導体の売り上げ (*handoudai no uriage*: sales of semiconductor)	3	39.9
乳製品の売り上げ (*nyuuseihin no uriage*: sales of dairy products)	2	28.2
海外での売り上げ (*kaigai deno uriage*: sales in foreign countries)	2	28.2
新製品の売り上げ (*sinseihin no uriage*: sales of new products)	4	50.8

Table 3 Examples of frequent expressions.

売り上げ (*uriage*: sales)	利益率の高い自社製品 (*riekiritu no takai zisya seihin*: products with high profit rate)
利益率 (*riekiritu*: profit rate)	販売価格の引き上げ (*hanbai kakaku no hikiage*: rise of price)
リストラ (*risutora*: restructuring)	高水準の受注残 (*kousuizyun no zyutyuuzan*: high level order backlog)

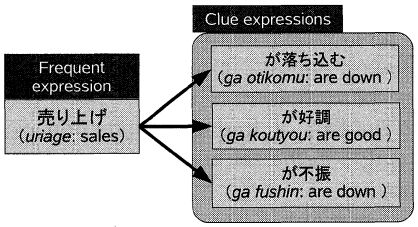

Fig. 3 Example of an appropriate frequent expression.

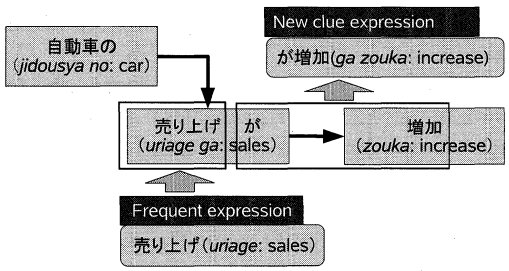

Fig. 4 Acquisition of new clue expressions.

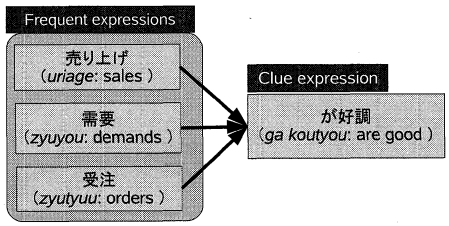

Fig. 5 Example of an appropriate clue expression.

entropy $H(e)$ larger than a threshold value calculated by Formula 9. Entropy $H(e)$ is used for reflecting "variety of clue expressions modified by frequent expression e". If entropy $H(e)$ is large, frequent expression e modifies various kinds of clue expressions and such a frequent expression is appropriate. (See Fig. 3.) Entropy $H(e)$ is calculated by the following Formula 7.

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s), \tag{7}$$

$$P(e,s) = \frac{f(e,s)}{\sum_{s' \in S(e)} f(e,s')}, \tag{8}$$

where,

$S(e)$**:** the set of clue expressions modified by frequent expression e.

$f(e,s)$**:** the number of frequent expression e's that modifies clue expression s in the set of articles concerning business performance.

The threshold value is calculated by the following Formula 9.

$$T_e = \alpha \log_2 |N_s|, \tag{9}$$

where, N_s is the set of clue expressions used for extracting frequent expressions and α is a constant ($0 < \alpha < 1$). Some examples of frequent expressions are shown in Table 3.

3.4 Acquisition of New Clue Expressions

The method for acquiring new clue expressions from frequent expressions is as follows.

Step 1: Extract a *bunsetu* modified by frequent expression e.

Step 2: Acquire new clue expression s by adding a case particle contained in the frequent expression e to the *bunsetu* (See Fig. 4.)[†].

Step 3: Calculate entropy $H(s)$ based on the probability $P(s,e)$ that clue expression s is modified by frequent expression e.

Step 4: Select clue expression s assigned entropy $H(s)$ larger than a threshold value calculated by the Formula 12.

Here, entropy $H(s)$ is introduced for selecting appropriate clue expressions and is calculated by the following Formula 10 (See Fig. 5.).

$$H(s) = -\sum_{e \in E(s)} P(s,e) \log_2 P(s,e), \tag{10}$$

[†]The reason why adding a case particle to the *bunsetu* of the clue expression is to be able to expect the improvement of precision of extracting causal expressions.

Table 4 Examples of clue expressions.

で補う (*de oginau*: cover)	が順調 (*ga zyuntyou*: go well)
で苦戦 (*de kusen*: make a poor fight)	が寄与 (*ga kiyo*: contribute)
が堅調 (*ga kentyou*: is robust)	が黒字転換した (*ga kurozi tenkan sita*: return to profitability)

$$P(s,e) = \frac{f(s,e)}{\sum_{e' \in E(s)} f(s,e')}, \tag{11}$$

where,

$E(s)$: the set of frequent expressions that modify clue expression s.

$f(s,e)$: the number of clue expression s's modified by frequent expression e in the set of articles concerning business performance.

The threshold value is calculated by the following Formula 12.

$$T_s = \alpha \log_2 |N_e|. \tag{12}$$

Here, N_e is a set of frequent expressions used for extracting clue expressions. α is the same constant that in Formula 9. Some examples of clue expressions are shown in Table 4.

3.5 Elimination of Inappropriate Clue Expressions and Frequent Expressions

Our method eliminates inappropriate clue expressions by using statistical information in the set of articles concerning business performance and the set of articles not concerning business performance. Note that we define the set of articles concerning business performance as Sa_p and the set of articles not concerning business performance as Sa_n. Moreover, the set of articles extracted by our method as articles concerning business performance is used as Sa_p and the set of articles not concerning business performance with the same number as Sa_p is used as Sa_n. Appropriate clue expressions frequently appear in Sa_p and inappropriate clue expressions frequently appear in not only Sa_p but also Sa_n. For this reason, our method uses Sa_p and Sa_n for eliminating inappropriate clue expressions. For example, although appropriate clue expression "が好調 (*ga koutyou:* were good)" frequently appears in Sa_p, inappropriate clue expression "こと (*koto:* thing)" frequently appears in not only Sa_p but also Sa_n.

The method for eliminating inappropriate clue expressions is as follows. First, our method calculates score $W(s, Sa_p)$ of clue expression s in Sa_p and score $W(s, Sa_n)$ of clue expression s in Sa_n by the following Formula 13.

$$W(s, Sa_p) = P(s, Sa_p) H(s, Sa_p), \tag{13}$$

where,

$P(s, Sa_p)$: the probability that a sentence containing clue expression s appears in Sa_p,

$H(s, Sa_p)$: the entropy based on the probability $P(s,d)$ that a sentence containing s appears in document $d \in Sa_p$.

$H(s, Sa_p)$ is calculated by the following formula 14.

$$H(s, Sa_p) = -\sum_{d \in Sa_p} P(s,d) \log_2 P(s,d) \tag{14}$$

$$P(s,d) = \frac{f(s,d)}{\sum_{d' \in Sa_p} f(s,d')} \tag{15}$$

Here, $f(s,d)$ is a number of sentences containing clue expression s in document d.

Next, our method compares $W(s, Sa_p)$ with $W(s, Sa_n)$. If score $W(s, Sa_p)$ is smaller than $2W(s, Sa_n)$, clue expression s is eliminated as an inappropriate clue expression. For example, since "こと (*koto:* thing)" frequently appears in not only Sa_p but also Sa_n, large $W(s, Sa_p)$ and $W(s, Sa_n)$ are assigned. Hence, "こと (*koto:* thing)" is eliminated as an inappropriate clue expression. Moreover, inappropriate frequent expressions are also eliminated by applying this method to frequent expressions.

Here, clue expressions and frequent expressions containing numbers are also eliminated to prevent extracting sale proceeds as a causal expression.

3.6 Extraction of Causal Expressions by Using Frequent Expressions and Clue Expressions

Finally, our method extracts causal expressions by using frequent expressions and clue expressions. A causal expression consists of a clue expression and a phrase that modifies the clue expression. Moreover, the phrase that modifies the clue expression contains some frequent expressions. For example, "新型の自動車の売り上げが好調 (*singata no jidousya no uriage ga koutyou*: sales of new car are good)" is a causal expression since phrase "新型の自動車の売り上げ (*singata no jidousya no uriage ga koutyou*: sales of new car)" modifies clue expression "が好調 (*ga koutyou*: are good)" and the phrase contains frequent expression "売り上げ (*uriage*: sales)".

Here, the same phrase may modify different clue expressions. For example, sentence "新型の自動車の売り上げが好調だった (*singata no jidousya no uriage ga koutyou datta* : sales of new car were good)" contains three clue expressions, "が好調 (*ga koutyou*: are good)", "が好調だ (*ga koutyou da*: are good)" and "が好調だった (*ga koutyou datta*: were good)". Hence, three causal expressions implying the same content are extracted in this case. In order to prevent it, if the same phrase modifies some clue expressions in a sentence, our method extracts a causal expression by adding the phrase to a clue expression assigned the largest entropy $H(s)$ among clue expressions modified by the phrase. In this case, since the largest entropy $H(s)$ is assigned to clue expression "が好調だった (*ga koutyou*

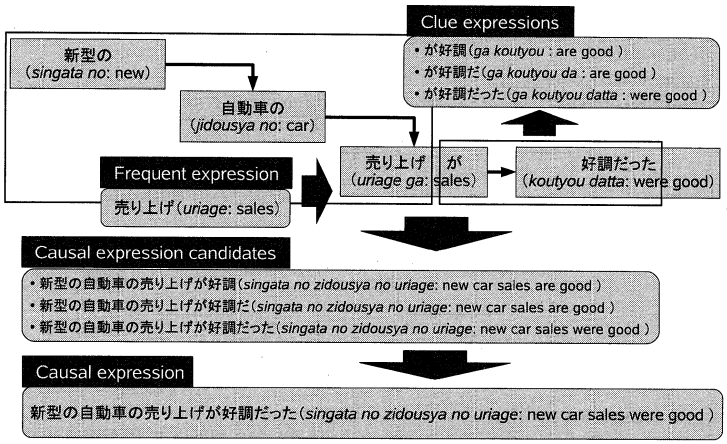

Fig. 6 Extraction of causal expressions.

Table 5 Examples of causal expressions extracted by our method.

causal expression	システム構築などソフト・サービス部門の収益が寄与する *sisutemukoutiku nado sofuto sarbis bumon no syuueki ga kiyo suru*: Profit of software and service sector including system construction contributes.
frequent expression	部門 (*bumon*: sector), 収益 (*syuueki*: profit)
clue expression	が寄与する (*ga kiyosuru*: contribute)
causal expression	情報関連部材などの落ち込みや繊維部門の不振が響く *zyouhou kanren buzai nado no otikomi ya senni bumon no hushin ga hibiku*: Slackness of information-related materials and textile sector affects the business performance.
frequent expression	落ち込み (*otikomi*:), 不振 (*hushin*: down), 部門 (*bumon*: sector)
clue expression	が響く (*hibiku*: affect)
causal expression	国内のリストラクチャリング（事業の再構築）が奏功 *kokunai no zigyou no saikoutiku ga soukou*: Domestic restructuring (restructuring of the business) succeeded.
frequent expression	リストラ (*risutora*: restructuring)
clue expression	が奏功 (*ga soukou*: succeed)

datta: were good)", the causal expression is "新型の自動車の売り上げが好調だった (*singata no jidousya no uriage ga koutyou datta* : sales of new car were good)". (See Fig. 6.)

4. Evaluation

4.1 Implementation

We implemented our method. Our method extracted 20,880 newspaper articles concerning business performance from Nikkei newspapers published from 2001 to 2005 and extracted causal expressions from them. Note that although articles not concerning business performance may be contained in the extracted 20,880 newspaper articles, they are not eliminated by hand. The initial clue expressions are "が好調 (*ga koutyou:* were good)" and "が不振 (*ga husin:* were down)". We employ ChaSen [†] as a Japanese morphological analyzer, and CaboCha [††] as a Japanese parser and SVM^{light} [†††] as an implementation of SVM. Some examples of causal expressions extracted by our method are shown in Table 5.

4.2 Precision and Recall

First, we evaluated our method for extracting articles concerning business performance. We manually selected 1,136 articles concerning business performance from Nikkei newspapers published from 2001 to 2005 as a correct data set, and calculated precision and recall [††††]. As a result, our method attained 93.7% recall and 88.6% precision, respectively.

Next, we evaluated our method for extracting causal expressions. We manually extracted 559 causal expressions from 131 articles concerning business performance as a correct data set. Moreover, we extracted causal expressions by our method from the same 131 articles as test data and calculated precision and recall. A causal expression extracted as the correct data is added information on a document and a sentence containing the causal expression. A causal expression extracted by our method is correct if the following two conditions are satisfied.

- A causal expression extracted by our method contains a causal expression extracted as the correct data.
- A document and a sentence containing a causal expres-

[†]http://chasen.aist-nara.ac.jp/hiki/ChaSen/
[††]http://chasen.org/~taku/software/cabocha/
[†††]http://svmlight.joachims.org
[††††]Note that Nikkei newspapers published in 2000 are not used as correct data since they are used as training data.

Table 6 Recall and precision of causal expression acquisition (T = 5).

α	num. of causal expressions	num. of f. exp.	num. of c. exp.	Precision (%)	Recall (%)
0.5	816	28	13	85.7	1.24
0.45	845	28	17	85.7	1.24
0.4	12250	86	139	79.3	13.0
0.35	40760	422	354	78.1	38.9
0.3	89963	850	922	79.2	66.1
0.25	138678	1424	1747	72.7	77.6
0.2	178102	2765	3381	65.9	80.8

Table 7 Recall and precision of causal expression acquisition (α = 0.3).

T	num. of causal expressions	num. of f. exp.	num. of c. exp.	Precision (%)	Recall (%)
1	3047	27	96	84.2	2.66
2	20429	248	246	78.5	21.0
3	51001	433	506	80.7	46.0
4	84334	665	812	80.4	64.1
5	89963	850	922	79.2	66.1
6	92447	884	938	78.4	67.1
7	92592	894	945	78.4	67.1
8	92788	899	950	78.4	67.1

Table 8 Comparison between our method and previous method.

	num. of causal expressions	num. of f. exp.	num. of c. exp.	Precision (%)	Recall (%)
our new method	89963	850	922	79.2	66.1
our previous method	104985	475	938	71.5	64.4

sion extracted by our method correspond to a document and a sentence containing a causal expression extracted as the correct data.

The precision and recall are calculated by the following formulas.

$$Precision = \frac{|E_{ce} \cap C_{ce}|}{|E_{ce}|},\ Recall = \frac{|E_{ce} \cap C_{ce}|}{|C_{ce}|},$$

where,

E_{ce}: the set of causal expressions extracted by our method from 131 articles concerning business performance used for making the correct data set.

C_{ce}: the set of causal expressions contained in the correct data set.

Note that $|E_{ce}|$ and $|C_{ce}|$ are respectively total numbers of causal expressions extracted by our method and extracted by hand from 131 articles concerning business performance used for making the correct data set.

Tables 6 and 7 show the results. Here, α is a parameter used for determining a threshold value in Formulas 9 and 12. T is a number of times in which frequent expressions and clue expressions are extracted. "num. of causal expressions" is a number of causal expressions extracted by our method from 20,880 newspaper articles concerning business performance. "f. exp." and "c. exp." denote "frequent expressions" and "clue expressions", respectively.

4.3 Comparison of Our Method

For confirming the effectiveness of our method, we compared our method with our previous one [2], which was originally developed for extracting expressions concerning traffic accident causes. Note that since our previous method was originally developed for extracting expressions concerning traffic accident causes, expressions to be extracted are different. Our previous method extracts expressions, e.g., "前方不注意 (*zenpou hutyuui*: not looking ahead carefully)" or "スピードの出し過ぎ (*supiid no dasisugi*: excessive speed)" as expressions implying traffic accident causes from articles concerning traffic accident. Although our method extracts causal expressions by using both frequent expressions and clue expressions, our previous method extracts expressions implying traffic accident causes by using only frequent expressions. Hence, the method for extracting expressions to be extracted from a document set is also different. Moreover, our new method improves our previous one by appending modules for eliminating inappropriate clue expressions and inappropriate frequent expressions. However, since our previous method is also able to extract frequent expressions and clue expressions for extracting causal expressions by changing initial clue expressions, we use our previous method as a comparison method. Here, we adjusted threshold values so that our new method and our previous method have the same recall. Table 8 shows the results.

5. Discussion

We consider that our method is able to extract causal expressions appropriately, since our method achieved 79.2% precision and 66.1% recall, respectively. Since a causal expression consists of a clue expression and frequent expressions, it is necessary not only to extract frequent expressions and

Table 9 Examples of inappropriate causal expressions extracted by our method.

causal expression	基礎化学品や汎用樹脂の採算が *kiso kagakuhin ya hanyou zyusi no saisann ga* The profit margin of basic chemical products and commodity plastics
frequent expression	化学品 (*kagakuhin*: chemical products)
clue expression	の採算が (*no saisan ga*: profit margin)
causal expression	ゲームソフトの売れ行きが *geimu sofuto no ureyuki ga* Sales of game software
frequent expression	ゲームソフト (*gaimu sofuto*: game software)
clue expression	の売れ行きが (*no ureyuki ga*: sales)

Table 10 Examples of causal expressions not extracted by our method.

causal expression	「真・三国無双 2」がヒット *"shin sangoku musou 2 ga hitto"* The "Dynasty Warriors 2" was hit
causal expression	「クラリーノ」の低迷 *"kurariino" no teimei* Depression of "Clarino"

clue expressions effective for extracting causal expressions but also not to extract inappropriate frequent expressions and clue expressions for achieving high precision. Some examples of inappropriate causal expressions extracted by our method are shown in Table 9. For example, "ゲームソフトの売れ行きが (*geimu sofuto no ureyuki ga*: Sales of game software)" is not a causal expression since it is not able to judge whether sales is "good" or "down". In this case, since "売れ行き (*ureyuki ga*: sales)", which should be acquired as a frequent expression, is acquired as a clue expression, it is extracted as a causal expression.

However, the recall is lower than the precision. Our method does not select a clue expression assigned small entropy $H(s)$ and a frequent expression assigned small entropy $H(e)$. Moreover, our method eliminates clue expressions and frequent expressions which seldom appear in the set of articles concerning business performance even if they are effective for extracting causal expressions. Many causal expressions not extracted by our method were ones that a product name modifies a clue expression. Some examples are shown in Table 10. Our method selects a frequent expression modifying various kinds of clue expressions as an appropriate one. Hence, it is necessary for appropriate frequent expressions to be contained in a number of documents concerning business performance of many companies. However, since the product names are contained only in documents concerning business performance of one company, they are not selected as appropriate frequent expressions. Since clue expressions and frequent expressions extracted by our method need to appear frequently in the set of articles concerning business performance, we consider that our method did not achieve high recall.

Experimental results shown in Table 8 suggest that our new method outperforms our previous one[†]. Our new method can effectively eliminate inappropriate clue expressions than our previous one due to improvement of process for eliminating inappropriate clue expressions. For this reason, our new method outperforms our previous one. Our new method and our previous one execute the step of acquiring clue expressions and the step of acquiring frequent expressions, iteratively. If many inappropriate clue expressions are included in the set of clue expressions for acquiring frequent expressions, many inappropriate frequent expressions may be acquired. Hence, the process for eliminating inappropriate clue expressions is essential for improving the performance. For example, "になる (*ni naru*: become)", which is an inappropriate clue expression, is acquired as a clue expression by our previous method. However, it is not acquired by our new method. "になる (*ni naru*: become)" is contained in not only the set of articles concerning business performance but also the set of articles not concerning business performance. Hence, it is not acquired by the method shown in Sect. 3.5, which is introduced only into our new method.

We consider that causal expressions extracted by our method are able to be used as data for computer trading. In the case of applying causal expressions to computer trading, we consider that precision is more important than recall. If expressions inappropriate as cause information are extracted as causal expression, there is a possibility with a money loss when they are used as data for computer trading. From the evaluation results, we consider that causal expressions are useful as data for computer trading since our method attained 79.2% precision. Moreover, high recall is able to be attained, if necessary, by assigning a low value to a threshold value (e.g. $\alpha = 0.2$). However, some technical problems still remain to be solved in using causal expressions as data for computer trading. For example, it is necessary to add polarity to causal expressions automatically (e.g., "car sales are good" is positive and "car sale are down" is negative). We consider that this can be solved if the polarity can be added to clue expressions automatically (i.e., "be good" is

[†]We confirmed, by statistical test, that the precision improved significantly.

positive and "be down" is negative.).

Although we adopt cause information of the business performance as target information to be extracted, we consider that various kinds of important information for the investment exist such as information concerning business tie-up or company bankruptcy. We expect that our method is able to extract causes information of business tie-up or company bankruptcy from them by using appropriate initial clue expressions for extracting them.

6. Related Work

As related work for extracting phrases that have a particular meaning, Riloff et al. proposed a method for learning extraction patterns for subjective expressions by applying syntactic templates made by hand to the training corpus [3]. Kanayama et al. proposed a method for extracting a set of sentiment units by using transfer-based machine translation engine replacing the translation patterns with sentiment patterns [4]. Kobayashi et al. proposed a semi-automatic method for collecting evaluative expressions by using particular co-occurrence patterns of evaluated subjects, focused attributes and values [5]. Morinaga et al. proposed a method for collecting and analyzing people's opinions about target products from Web pages by using an evaluation-expression dictionary and syntactic property rules learned manually from training examples [6]. However, to construct a complete list of complex rules or patterns manually, which is the case of the above methods, is a time-consuming and costly task. Moreover, the rules and the patterns made by hand may be domain-specific and can not be applied to other tasks. In contract, our method uses statistical information and only two initial clue expressions consisting of two words as an initial input. The domain-specific dictionaries, predetermined patterns, complex rules made by hand are not needed. Hence, our method is expected to be applied to other tasks for acquiring phrases that have a particular meaning not limited to cause information (e.g. opinion information, reputation information) by changing an initial input.

Inui et al. proposed a method for acquiring causal relations (*cause*, *effect*, *precond* and *means*) from a complex sentence containing a Japanese resultative connective ため (*tame*) [7]. Khoo et al. proposed a method for extracting cause-effect information from a newspaper text and a method for extracting causal knowledge from a medical database by applying patterns made by hand [8], [9]. In these researches, *cause* and *effect* etc. need to be contained together in a sentence. In our method, *cause* corresponds to causal expressions and *effect* corresponds to business performance. Here, our method is able to extract causal expressions from a sentence contained in an article concerning business performance. Hence, our method need not assume that *cause* and *effect* are contained together in a sentence.

Koppel et al. proposed a method for classifying news stories about a company according to its apparent impact on the performance of the company's stock [10]. Lavrenko et al. proposed a method for identifying news stories that influence the behavior of financial markets [11]. In contrast, since our method extracts causal expressions from newspaper articles concerning business performance, the task is different. In general, articles concerning business performance contain content that influences the stock price. However, even if the business performance of a company is good, the stock price of the company will not rise if the main cause is not related to its core business (e.g. profit from sales of stocks). Hence, we consider that it is necessary not only to classify articles whether they influence the stock prise but also to analyze content of articles. We expect to be able to analyze content of the articles concerning business performance circumstantially by using causal expressions extracted by our method.

7. Conclusion

We proposed a method for extracting phrases implying cause information from Japanese financial articles concerning business performance. First, our method extracts articles concerning business performance from newspaper corpus. Next, our method extracts causal expressions from them by using statistical information and initial clue expressions. We evaluated our method and it attained 79.2% precision and 66.1% recall, respectively. In addition to this, we compared our method with our previous method [2] by experiments and the experimental results showed that our new method outperforms our previous one.

Acknowledgment

We express our gratitude to Nikkei Shimbun, Inc. who permits us to use the Nikkei news paper articles in a machine readable form. This work was supported in part by Grant-in-Aid for Priority Areas (B)(2) 16092213 and the Global COE Program (Frontiers of Intelligent Sensing) from The Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] V. Vapnik, Statistical Learning Theory, Wiley, 1999.
[2] H. Sakai, S. Umemura, and S. Masuyama, "Extraction of expressions concerning accident cause contained in articles on traffic accidents," Journal of Natural Language Processing, vol.13, no.4, pp.99–124, 2006.
[3] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," Proc. 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.105–112, 2003.
[4] H. Kanayama, T. Nasukawa, and H. Watanabe, "Deeper sentiment analysis using machine translation technology," Proc. 20th COLING, pp.494–500, 2004.
[5] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and Y. Fukushima, "Collecting evaluative expressions for opinion extraction," Journal of Natural Language Processing, vol.12, no.3, pp.203–222, 2005.
[6] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," Proc. Eighth ACM SIGKDD Int. Conf. on KDD2002, pp.341–349, 2002.

[7] T. Inui, K. Inui, and Y. Matsumoto, "Acquiring causal knowledge from text using the connective marker *tame*," J. IPSJ, vol.45, no.3, pp.919–933, 2004.
[8] C.S. Khoo, J. Kornfilt, R.N. Oddy, and S.H. Myaeng, "Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing," Literary and Linguistic Computing, vol.13, no.4, pp.177–186, 1998.
[9] C.S. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," Proc. 38th ACL, pp.336–343, 2000.
[10] M. Koppel and I. Shtrimberg, "Good news or bad news? let the market decide," Proc. AAAI Spring Symposium on Exploring Attitude and Affect in Text, pp.86–88, 2004.
[11] V. Lavrenko, M. Schmill, D. Lawrie, and P. Ogilvie, "Mining of concurrent text and time series," Proc. KDD 2000 Conference Text Mining Workshop, pp.37–44, 2000.

Hiroyuki Sakai is presently an Assistant professor at the Department of Knowledge-based Information Engineering, Toyohashi University of Technology. He received the B.E., M.E. and D.E. degrees in Engineering from Toyohashi University of Technology, in 2000, 2002 and 2005, respectively. His research interest centers around natural language processing including automatic text summarization, knowledge acquisition from corpus and information retrieval. Dr. Sakai is a member of the Japanese Society for Artificial Intelligence, the Association for Natural Language Processing of Japan.

Shigeru Masuyama is presently a Professor at the Department of Knowledge-Based Information Engineering, Toyohashi University of Technology. He received the B.E., M.E. and D.E. degrees in Engineering (Applied Mathematics and Physics) from Kyoto University, in 1977, 1979 and 1983, respectively. He was with the Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University from 1984 to 1989. He joined the Department of Knowledge-Based Information Engineering, Toyohashi University of Technology in 1989. His research interest includes computational graph theory and natural language processing, e,g., automatic text summarization and knowledge acquisition from corpus. Dr. Masuyama is a member of the OR society of Japan, the Information Processing Society of Japan, the Institute of Systems, Control, Information Engineers of Japan and the Association for Natural Language Processing of Japan, etc.