# Extraction of Cause Information from Newspaper Articles Concerning Business Performance

Hiroyuki Sakai[1] and Shigeru Masuyama[1]

Department of Knowledge-based Information Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
(sakai,masuyama)@smlab.tutkie.tut.ac.jp

**Abstract.** We propose a method of extracting cause information from Japanese newspaper articles concerning business performance. 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 phrases 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 or language for acquiring phrases that have a particular meaning not limited to cause information. We compared our method with our previous method originally proposed for extracting phrases concerning traffic accident causes and experimental results showed that our new method outperforms our previous one.

## 1 Introduction

We propose a method of extracting cause information from Japanese newspaper 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”. Cause information is useful for investors in selecting companies to invest. 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. Moreover, cause information of the business performance is also important, because, even if the business performance is good, the 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. However, since there are a number of companies that announce business performance, acquiring their all cause information manually is a considerably hard task. 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 2 initial clue phrases. Here, the “clue phrases” are defined as phrases frequently modified by causal expressions.

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[5]. 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[2][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[1]. 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[4]. 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 patters made by hand may be domain-specific and can not be applied to other tasks. In contract, our method uses statistical information and only 2 initial clue phrases consisting of 2 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 or language for acquiring phrases that have a particular meaning not limited to cause information (e.g. opinion information, reputation information) by changing an initial input. We preliminarily proposed a method for extracting phrases concerning traffic accident causes by using statistical information and initial clue phrases[6]. However, our previous method could not attain high precision if inappropriate phrases are extracted. In this paper, we also propose a method for eliminating inappropriate phrases automatically and, by introducing it, our new method attains high precision.

## 2 Extraction of articles concerning business performance

As a preparation, our method extracts articles concerning business performance from newspaper corpus by using Support Vector Machine (SVM) [7].

### 2.1 Feature selection

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 $H(t_i, S_p)$ 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 2:

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d),\ P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d \in S_p} tf(t_i, d)}, \tag{2}$$

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 extracted as a feature of SVM. By introducing Entropy $H(t_i, S_p)$, a large score is assigned 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 extracted as a feature. 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.

## 3 Extraction of causal expressions

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

### 3.1 Acquisition of clue phrases

Our method for acquiring clue phrases is as follows.

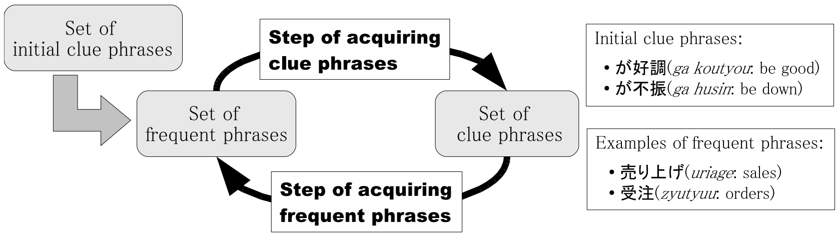

**Fig. 1.** Outline of our method

Step 1: Input some initial clue phrases and acquire phrases that modify them. Here, we use two clue phrases, "が好調 (*ga koutyou*: be good)" and "が不振 (*ga husin*: be down)", as initial clue phrases.
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 phrases extracted in Step 2 are defined as "frequent phrases".
Step 3: Acquire new clue phrases modified by the frequent phrases.
Step 4: Extract new frequent phrases from a set of phrases that modify the new clue phrases acquired in Step 3. This step is the same as Step 2.
Step 5: Repeat Steps 3 and 4 until predetermined times or neither new clue phrases nor new frequent phrases are extracted.

An outline of the method is illustrated in Figure 1.

### 3.2 Extraction of Frequent phrases

The method for extracting "frequent phrases" from a set of phrases that modify clue phrases is as follows.

Step 1: Acquire a *bunsetu* modifying a clue phrase and eliminate a case particle from the *bunsetu*. Here, the *bunsetu* is defined as $c$.
Step 2: Acquire frequent phrase candidates by adding *bunsetu* modifying $c$ to $c$. (See Figure 2.)
Step 3: Calculate score $S_f(e, c)$ of frequent phrase candidate $e$ containing $c$ by the following Formula 3.
Step 4: Adopt $e$ assigned the best score $S_f(e, c)$ in the set of frequent phrase candidates containing $c$ as the frequent phrase.

Score $S_f(e, c)$ is calculated by the following Formula 3:

$$S_f(e, c) = -f_e(e, c)\sqrt{f_p(e)}\log_2 P(e, c), \qquad (3)$$

where, $P(e, c)$ is the probability that frequent phrase candidate $e$ containing $c$ appears in the set of articles concerning business performance. $f_e(e, c)$ is the number of frequent phrase candidate $e$'s containing $c$. $f_p(e)$ is the number of *bunsetu*'s that compose $e$. Here, $P(e, c) = f_e(e, c)/Ne(c)$, where, $Ne(c)$ is the total number of frequent phrase candidates containing $c$ in the set of articles concerning business performance.

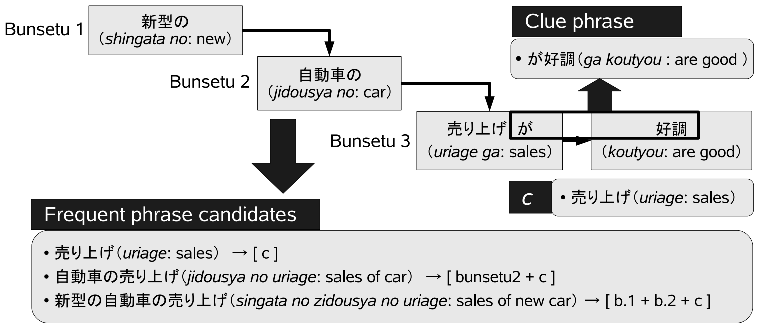
Bunsetu 1
新型の （shingata no: new）
Bunsetu 2
自動車の （jidousya no: car）
Bunsetu 3
売り上げ が （uriage ga: sales）
好調 （koutyou: are good）
Clue phrase
• が好調（ga koutyou : are good ）
c
• 売り上げ（uriage: sales）
Frequent phrase candidates
• 売り上げ（uriage: sales） → [ c ]
• 自動車の売り上げ（jidousya no uriage: sales of car） → [ bunsetu2 + c ]
• 新型の自動車の売り上げ（singata no zidousya no uriage: sales of new car） → [ b.1 + b.2 + c ]

**Fig. 2.** Examples of frequent phrase candidates

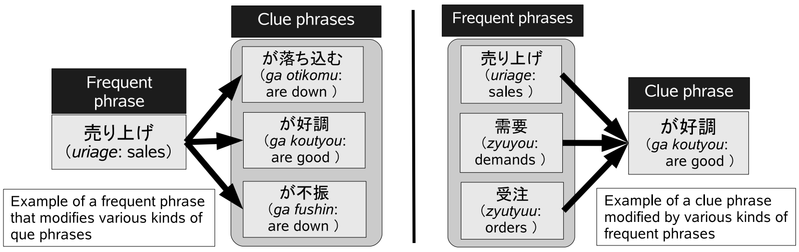
Frequent phrase
売り上げ （uriage: sales）
Example of a frequent phrase that modifies various kinds of que phrases
Clue phrases
が落ち込む （ga otikomu: are down ）
が好調 （ga koutyou: are good ）
が不振 （ga fushin: are down ）
Frequent phrases
売り上げ （uriage: sales ）
需要 （zyuyou: demands ）
受注 （zyutyuu: orders ）
Clue phrase
が好調 （ga koutyou: are good ）
Example of a clue phrase modified by various kinds of frequent phrases

**Fig. 3.** Example of an appropriate frequent phrase and an appropriate que phrase

### 3.3 Selection of frequent phrases

The frequent phrases extracted from a set of phrases that modify clue phrases may contain inappropriate ones. Hence, our method selects appropriate frequent phrases from them. Here, our method calculates entropy $H(e)$ based on the probability $P(e, s)$ that frequent phrase $e$ modifies clue phrase $s$ and selects frequent phrases assigned entropy $H(e)$ larger than a threshold value. Entropy $H(e)$ is used for reflecting "variety of clue phrases modified by frequent phrase $e$". If entropy $H(e)$ is large, frequent phrase $e$ modifies various kinds of clue phrases and such a frequent phrase is appropriate. (See Figure 3.) The entropy $H(e)$ is calculated by the following Formula 4.

$$H(e) = -\sum_{s \in S(e)} P(e, s) \log_2 P(e, s), \;\; P(e, s) = \frac{f(e, s)}{\sum_{s \in S(e)} f(e, s)}, \tag{4}$$

where, $S(e)$ is the set of clue phrases modified by frequent phrase $e$ and $f(e, s)$ is the number of frequent phrases $e$'s that modifies clue phrase $s$ in the set of articles concerning business performance. The threshold value is calculated by the following Formula 5.

$$T_e = \alpha \log_2 |N_s|, \tag{5}$$

where, $N_s$ is the set of clue phrases used for extracting frequent phrases and $\alpha$ is a constant $(0 < \alpha < 1)$.

### 3.4 Acquisition of clue phrases

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

Step 1: Extract a *bunsetu* modified by frequent phrase $e$.
Step 2: Acquire clue phrase $s$ by adding a case particle contained in the frequent phrase $e$ to the *bunsetu*.
Step 3: Calculate entropy $H(s)$ based on the probability $P(s, e)$ that clue phrase $s$ is modified by frequent phrase $e$.
Step 4: Select clue phrase $s$ assigned entropy $H(s)$ larger than a threshold value calculated by the Formula 5. (In this case, $N_s$ is a set of frequent phrases used for extracting clue phrases.)

Here, entropy $H(s)$ is introduced for selecting appropriate clue phrases and is calculated by the following Formula 6 (See Figure 3.).

$$H(s) = -\sum_{e \in E(s)} P(s, e) \log_2 P(s, e), \tag{6}$$

where, $P(s, e)$ is the probability that clue phrase $s$ is modified by frequent phrase $e$ and $E(s)$ is the set of frequent phrases that modify clue phrase $s$.

### 3.5 Elimination of inappropriate clue phrases

Finally, our method eliminates inappropriate clue phrases by using statistical information in the set of articles concerning business performance and the set of articles not concerning business performance. The method for eliminating inappropriate clue phrases is as follows. First, our method calculates score $W(s, S_p)$ of clue phrase $s$ in set $S_p$ of articles concerning business performance and score $W(s, S_n)$ of clue phrase $s$ in set $S_n$ of articles not concerning business performance by the following Formula 7.

$$W(s, S_p) = P(s, S_p)H(s, S_p), \tag{7}$$

where, $P(s, S_p)$ is the probability that a sentence containing clue phrase $s$ appears in $S_p$, and $H(s, S_p)$ is the entropy based on the probability $P(s, d)$ that a sentence containing $s$ appears in document $d \in S_p$. Next, our method compares $W(s, S_p)$ with $W(s, S_n)$. If score $W(s, S_p)$ is smaller than $2W(s, S_n)$, clue phrase $s$ is eliminated as an inappropriate clue phrase. Moreover, inappropriate frequent phrases are also eliminated by applying this method to frequent phrases. Here, clue phrases and frequent phrases containing numbers are also eliminated to prevent extracting sale proceeds as a causal expression.

## 4 Evaluation

We implemented our method. Our method extracted $20,880$ newspaper articles concerning business performance from Nikkei newspapers published from

**Table 1.** Recall and precision of causal expression acquisition

| $\alpha$ | num. of clue phrases | Precision (%) | Recall (%) |
|---|---:|---:|---:|
| 0.5 | 12 | 85.7 | 1.25 |
| 0.4 | 139 | 77.7 | 12.8 |
| 0.3 | 922 | 77.7 | 66.0 |
| 0.2 | 3381 | 63.7 | 80.1 |

**Table 2.** Comparison between our method and previous method

| | num. of clue phrases | Precision (%) | Recall (%) |
|---|---:|---:|---:|
| our new method | 922 | 77.7 | 66.0 |
| our previous method | 938 | 70.5 | 64.6 |

2001 to 2005 and extracted causal expressions from them. We employ ChaSen[1] as a Japanese morphological analyzer, and CaboCha[2] as a Japanese parser and $SVM^{light}$[3] as an implementation of SVM. 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 and calculated precision and recall. Here, a causal expression extracted by our method is correct if it contains a causal expression extracted as the correct data set. Table 1 shows the results. Here, $\alpha$ is a parameter used for determining a threshold value in Formula 5. For confirming the effectiveness of our method, we compared our method with our previous method[6], which was originally developed for extracting phrases concerning traffic accident causes. Note that our new method is a one that improves our previous method for eliminating inappropriate clue phrases and inappropriate frequent phrases. Table 2 shows the results.

## 5 Discussion

Experimental results shown in Table 2 suggest that our new method outperforms our previous method. The reason why our new method outperforms our previous method is that our new method can effectively eliminate inappropriate clue phrases than our previous method due to improvement of process for eliminating inappropriate clue phrases. Our new method and our previous

[1] http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[2] http://chasen.org/~taku/software/cabocha/

[3] http://svmlight.joachims.org

method process the step of acquiring clue phrases and the step of acquiring frequent phrases, iteratively. If many inappropriate clue phrases are included in the set of clue phrases for acquiring frequent phrases, many inappropriate frequent phrases may be acquired. Hence, the process for eliminating inappropriate clue phrases is important for improving the performance. For example, "になる (*ni naru*: become)", which is an inappropriate clue phrase, is acquired as a clue phrase 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 Section 3.5, which is introduced only into our new method.

## 6 Conclusion

We proposed a method for extracting phrases implying cause information from Japanese newspaper 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 phrases. We evaluated our method and it attained 77.7% precision and 66.0% recall, respectively. In addition to this, we compared our method with our previous method[6] by experiments and the experimental results showed that our new method outperforms our previous one.

## References

1. Kanayama, H., Nasukawa, T. and Watanabe, H.: Deeper sentiment analysis using machine translation technology, *Proceedings of the 20th COLING*, pp. 494–500 (2004).
2. Khoo, C. S., Kornfilt, J., Oddy, R. N. and Myaeng, S. H.: 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).
3. Khoo, C. S., Chan, S. and Niu, Y.: Extracting Causal Knowledge from a Medical Database Using Graphical Patterns, *Proceedings of the 38th ACL*, pp. 336–343 (2000).
4. Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T.: Mining product reputations on the Web, *Proceedings of Eighth ACM SIGKDD Int. Conf. on KDD2002*, pp. 341–349 (2002).
5. Riloff, E. and Wiebe, J.: Learning Extraction Patterns for Subjective Expressions, *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 105–112 (2003).
6. Sakai, H., Umemura, S. and Masuyama, S.: 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(in Japanese)).
7. Vapnik, V.: *Statistical Learning Theory*, Wiley (1999).