PAPER *Special Section on Natural Language Processing and its Applications*

Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies

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

SUMMARY We propose a method of assigning polarity to causal information extracted from Japanese financial articles concerning business performance of companies. Our method assigns polarity (positive or negative) to causal information in accordance with business performance, e.g. "*zidousya no uriage ga koutyou*: (Sales of cars are good)" (The polarity positive is assigned in this example). We may use causal expressions assigned polarity by our method, e.g., to analyze content of articles concerning business performance circumstantially. First, our method classifies articles concerning business performance into positive articles and negative articles. Using them, our method assigns polarity (positive or negative) to causal information extracted from the set of articles concerning business performance. Although our method needs training dataset for classifying articles concerning business performance into positive and negative ones, our method does not need a training dataset for assigning polarity to causal information. Hence, even if causal information not appearing in the training dataset for classifying articles concerning business performance into positive and negative ones exist, our method is able to assign it polarity by using statistical information of this classified sets of articles. We evaluated our method and confirmed that it attained 74.4% precision and 50.4% recall of assigning polarity positive, and 76.8% precision and 61.5% recall of assigning polarity negative, respectively.

key words: *polarity assignment, text mining, causal information, knowledge extraction, information extraction*

1. Introduction

Recently, natural language processing methods are expected to be used in financial applications. For example, a method that analyzes financial articles may provide useful information for supporting decision on investment. In particular, by the recent increase of the number of private investors in Japan, some methods for supporting them to make decision on investment are desired.

Collecting information concerning business performance of companies is a very important task for investment. If the business performance of a company is good, its stock price will rise in general. Moreover, causal information on the business performance is also important. For example, even if the business performance of a company is good, its stock price will not rise when the main cause of change in the business performance is the recording of an extraordinary profit not related to its core business (e.g. profit from sales of stocks). This is also the case for the bad business performance. Hence, causal 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 all of their causal information manually is a considerably hard and expensive task[*]. Hence, we proposed a method of identifying articles concerning business performance of companies and extracting causal information (e.g. "*zidousya no uriage ga koutyou*: Sales of cars are good") from them automatically[**] [1]. We defined a phrase implying causal information as a "causal expression".

In this paper, we propose a method of assigning polarity (either positive or negative) to causal expressions. We expect that causal expressions may be used as data, e.g., for computer trading and business trend forecast. However, in order to make causal expressions more effective information, it is desirable to assign them polarity (positive or negative) according to business performance. For example, polarity of causal expression "car sales are good" is positive and polarity of causal expression "car sales are down" is negative[***]. If the number of causal expressions that are assigned polarity positive increases, business condition is expected to recover. In contrast, if the number of causal expressions that are assigned polarity negative increases, business condition may slow down (e.g., see Fig. 4.). Moreover, we may analyze content of articles concerning business performance circumstantially by using causal expressions assigned polarity by our method. In addition, we are also able to provide causal expressions assigned polarity by our method as useful information for supporting decision on investment to private investors. Hence, we propose a method of assigning polarity (either positive or negative) to causal expressions automatically in this paper.

In our previous method [1], we defined causal expressions to be extracted as expressions that contain some "frequent expressions" and a "clue expression". Then, we defined a frequent expression as a phrase frequently appearing in a set of causal expressions and a clue expression as a phrase frequently modified by causal expressions. For example, in a causal expression "*zidousya no uriage ga*

Manuscript received March 20, 2009.
Manuscript revised July 17, 2009.

[†]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: 10.1587/transinf.E92.D.2341

[*] The number of companies listed on the Tokyo Stock Exchange is about 2397.

[**] We briefly introduce this method in Sect. 3.

[***] Note that polarity "neutral" is not used to assign to causal expressions since the causal expressions appear in articles concerning business performance only when the business performance changes.

koutyou: (Sales of cars are good)", a frequent expression is "*uriage*: (sales)" and a clue expression is "*ga koutyou*: (are good)". Our previous method acquired such frequent expressions and clue expressions automatically and extracted causal expressions by using these frequent expressions and clue expressions.

Our method assigns polarity to the causal expressions by using statistical information of the combination of frequent expressions and clue expressions. In general, polarity of causal expressions containing a clue expression "*ga koutyou*: (are good)" is positive. However, the polarity is not able to be determined only by clue expressions. For example, although the polarity of a causal expression "*zidousya no uriage ga zouka*: (car sales are increasing)" containing a clue expression "*ga zouka*: (are increasing)" is positive, the polarity of a causal expression "*risutora hiyou ga zouka*: (restructuring cost is increasing)" containing the same clue expression is negative. Hence, it is necessary for assigning polarity to employ the combination of frequent expressions and clue expressions. However, since the number of combinations of frequent expressions and a clue expression is enormous, it is impossible to assign polarity to causal expressions manually. Moreover, machine learning methods can assign polarity to only causal expressions that contain some combinations of frequent expressions and a clue expression that appear in training data.

In contrast, our method classifies articles concerning business performance into two categories, positive articles and negative articles, according to business performance. That is, our method classifies an article suggesting that business performance improves into the set of positive articles and classifies an article suggesting that business performance declines into the set of negative articles. After that, our method assigns polarity to a causal expression by using probability that combinations of frequent expressions and a clue expression appear in a set of positive articles or that of negative ones. For example, the combination of a frequent expression "*risutora hiyou*: (restructuring cost)" and a clue expression "*ga zouka*: (is increasing)" is frequently contained in a set of negative articles. Hence, our method successfully assigns polarity "negative" to causal expressions containing the combination of a frequent expression "*risutora hiyou*: (restructuring cost)" and a clue expression "*ga zouka*: (is increasing)".

We describe related work in Sect. 2 and explain difference between our proposed method and the related work. We briefly introduce our previous method to help understand our method of assigning polarity in Sect. 3. After that, we introduce our method in Sect. 4. Experimental results of evaluation are reported in Sect. 5 and analysis of the results on the experiments are discussed in Sect. 6. Section 7 concludes this paper.

2. Related Work

As related work, Takamura et al. proposed models for semantic orientations of phrases that consist of multiple words as well as classification methods based on the models by using a machine learning method [2]. They focused on "noun+adjective" and introduced latent variables into the model in order to capture the property of such phrases. Sakaji et al. proposed a method to automatically extract basis expressions that indicate economic trends from newspaper articles and a method to classify them into positive expressions that indicate upbeat, and negative expressions that indicate downturn in economy, by using a machine learning method [3]. However, these methods are not able to classify phrases that consist of words not appearing in the training dataset. In contrast, our method classifies articles concerning business performance into positive and negative ones. After that, our method assigns polarity to a causal expression by using probability that combinations of frequent expressions and a clue expression appear in a set of positive articles or that of negative ones. Although our method needs a training dataset for classifying articles concerning business performance into positive and negative ones, our method does not need a training dataset for assigning polarity to causal expressions. Hence, even if a causal expression not appearing in the training dataset for classifying articles exists, our method is able to assign it polarity by using statistical information of this classified sets of articles.

Turney proposed a method for classifying reviews as *recommended* (thumbs up) or *not recommended* (thumbs down) by calculating the mutual information between phrases in the review and a positive reference word "excellent", and a negative reference word "poor", respectively [4]. Wilson et al. proposed a method for recognizing contextual polarity in phrase-level sentiment analysis by using the BoosTexter AdaBoost.HM [5] machine learning algorithm and some features, e.g., polarity of words that compose the phrase [6]. (Here, the polarity of words was assigned by hand.) Baron et al. proposed a method for classifying collocations extracted by Xtract [7] into "positive" and " negative" [8]. The method classifies them by using the orientations of the words in the neighboring sentences. However, since the number of frequent expressions and clue expressions necessary to extract causal expressions is enormous, assigning polarities to frequent expressions and clue expressions manually is a considerably hard task. Moreover, the polarity is not able to be determined only by clue expressions, e.g., "*ga zouka*: (are increasing)". Hence, a method that uses the polarity of words that compose a causal expression, e.g., [4], [6], [8], is not applicable in our task.

Takamura et al. proposed a method for extracting semantic orientations of phrases (pairs of an adjective and a noun), which is applicable also to the pairs consisting of an adjective and an unseen noun [9]. They adopt the Potts model [10] for the probability model of the lexical network. However, a causal expression consists of a number of words and an adjective seldom appears in it (The probability that a causal expression containing an adjective appears is 0.09). Kaji et al. proposed a method for acquiring polar phrases (adjectives and "noun + post-positional particles + adjective") by using frequency in polar sentence corpus [11].

Here, the polar sentences are extracted by using lexicon-syntactic patterns and manually-created cue words list. In their method, the same polar phrases need to appear at least three times in both the set of positive polar sentences and that of negative ones. However, since a causal expression consists of a number of words, the same causal expression does not appear in the set of articles concerning business performance. Hence, Kaji et al.'s method is not applicable to our task. In contrast, our method solves the problem that the same causal expression does not appear by replacing a causal expression that consists of a number of words with some "frequent expressions" and a "clue expression".

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 [12]. Lavrenko et al. proposed a method for identifying news stories that influence the behavior of financial markets [13]. In contrast, since our method assigns polarity to causal expressions extracted from newspaper articles concerning business performance, the task is different. In general, articles concerning business performance contain some 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. Hence, we consider that it is necessary not only to classify articles whether they influence the stock price 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 assigned polarity by our method.

3. Extraction of Causal Expressions

We proposed a method of extracting causal expressions from financial articles concerning business performance automatically [1]. In this section, we briefly introduce our previous method to help understand our method of assigning polarity. We defined causal expressions to be extracted as expressions that contain some "frequent expressions" and one "clue expression", and our previous method extracts causal expressions by acquiring the frequent expressions and the clue expressions automatically. The method for acquiring frequent expressions and clue expressions is as follows.

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

Step 2: Extract phrases that frequently appear in a set of the phrases acquired in Step 1 as frequent expressions. Here, the phrases acquired in Step 1 are defined as frequent expression candidates and appropriate frequent expressions are selected from them.

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 our previous method is shown in Fig. 1.

3.1 Selection of Frequent Expressions

Our previous method [1] selects appropriate frequent expressions from a set of frequent expression candidates. Here, our previous method calculates entropy $H(e)$ based on the probability $P(e, s)$ that frequent expression e modifies clue expression s and selects a frequent expression that is assigned entropy $H(e)$ larger than a threshold value calculated by Formula 2. 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. 2.) Entropy $H(e)$ is calculated by the following Formula 1:

$$H(e) = -\sum_{s \in S(e)} P(e, s) \log_2 P(e, s). \tag{1}$$

Here, $S(e)$ is the set of clue expressions modified by frequent expression e in the set of articles concerning business

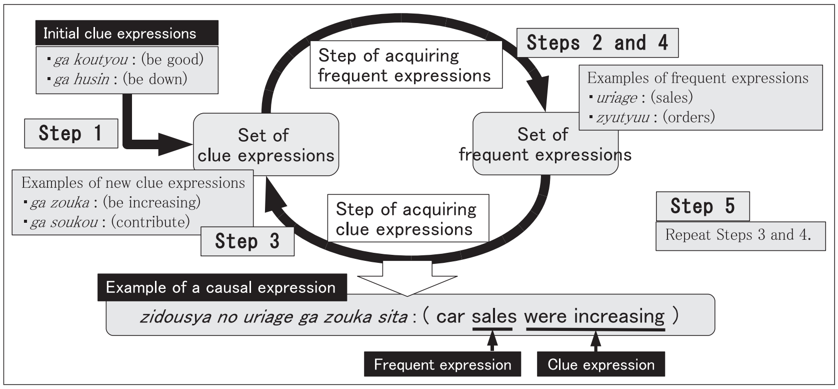
Initial clue expressions
・ga koutyou : (be good)
・ga husin : (be down)
Step 1
Set of clue expressions
Step of acquiring frequent expressions
Steps 2 and 4
Examples of frequent expressions
・uriage : (sales)
・zyutyuu : (orders)
Set of frequent expressions
Examples of new clue expressions
・ga zouka : (be increasing)
・ga soukou : (contribute)
Step 3
Step of acquiring clue expressions
Step 5
Repeat Steps 3 and 4.
Example of a causal expression
zidousya no uriage ga zouka sita : (car sales were increasing)
Frequent expression
Clue expression

Fig. 1 Outline of our previous method.

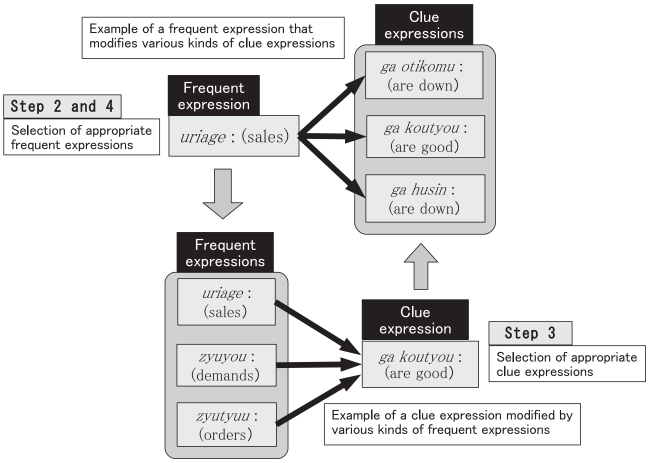

Fig. 2 Example of an appropriate frequent expression and an appropriate clue expression.

performance. The threshold value is calculated by the following Formula 2:

$$T_e = \alpha \log_2 |N_s|. \tag{2}$$

Here, N_s is the set of clue expressions and α is a constant ($0 < \alpha < 1$).

3.2 Acquisition of New Clue Expressions

Our previous method [1] acquires new clue expressions by using frequent expressions. First, our previous method extracts a *bunsetu*[†] modified by a phrase containing frequent expression e and acquires new clue expression s by adding a case particle contained in the phrase to the *bunsetu*. Next, our previous method calculates entropy $H(s)$ based on the probability $P(s, e)$ that clue expression s is modified by frequent expression e and selects clue expression s that is assigned entropy $H(s)$ larger than a threshold value calculated by Formula 4 to be introduced hereafter:

$$H(s) = -\sum_{e \in E(s)} P(s, e) \log_2 P(s, e). \tag{3}$$

Here, $E(s)$ is the set of frequent expressions that modify clue expression s. The threshold value is calculated by the following Formula 4:

$$T_s = \alpha \log_2 |N_e|. \tag{4}$$

Here, N_e is the set of frequent expressions and α is the same constant value that in Formula 2.

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

Finally, our previous 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 (see Fig. 1.). For example, "*tyuugoku muke no ekisyou kanren no densi buhin ga kaihuku suru*: (The amount of electronic parts related to liquid-crystal exported for China recover)" is extracted as a causal expression since phrase "*tyuugoku muke no ekisyou kanren no densi buhin*: (The amount of electronic parts related to liquid-crystal exported for China)" modifies clue expression "*ga kaihuku suru*: (recover)" and the phrase contains two frequent expressions "*densi buhin*: (electronic parts)" and "*ekisyou kanren*: (related to liquid-crystal)".

4. Polarity Assignment to Causal Expressions

Our method assigns polarity to causal expressions by using statistical information of combinations of "frequent expressions" and a "clue expression". For example, the combinations of a frequent expression "*uriage*: (sales)" and a clue expression "*ga zouka*: (are increasing)" is frequently contained in a set of articles suggesting that business performance improves. In contrast, the combinations of a frequent expression "*risutora hiyou*: (restructuring cost)" and a clue expression "*ga zouka*: (are increasing)" is frequently contained in a set of articles suggesting that business performance deteriorates. Hence, our method classifies articles concerning business performance into two categories, positive articles and negative articles, according to business performance. That is, our method classifies an article suggesting that business performance improves into positive articles and classifies an article suggesting that business performance declines into negative articles. After that, our method assigns polarity to a causal expression by using probability that combinations of frequent expressions and a clue expression appear in a set of positive articles or that of negative ones. An outline of our method is shown in Fig. 3:

4.1 Classification of Articles Concerning Business Performance

We explain about the method of classifying articles concerning business performance in this subsection. Our method classifies articles into positive and negative ones by using Support Vector Machine (SVM). First, we extract articles concerning business performance from a set of financial articles and manually classify them into positive and negative ones. They are used as a training data of SVM. Note that the training data is used to classify the articles into positive and negative ones, and is not used to assign polarity to causal expressions. Next, our method extracts content words effective for classifying articles as features from the training data. Actually, our method calculates score $W(t_i, S_p)$ of content word t_i contained in positive articles set S_p and score $W(t_i, S_n)$ of content word t_i contained in negative articles set S_n by the following Formula 5:

$$W(t_i, S_p) = P(t_i, S_p) H(t_i, S_p). \tag{5}$$

Here, $P(t_i, S_p)$ is the probability that word t_i appears in S_p

[†] A *bunsetu* is a basic block in Japanese composed of several words.

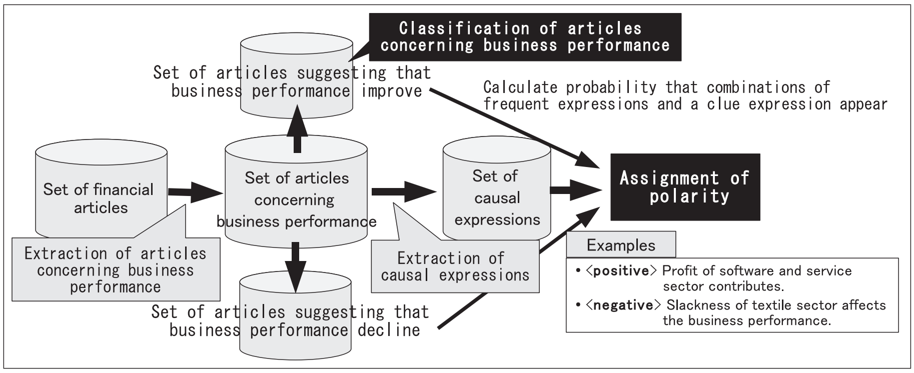
Classification of articles concerning business performance
Set of articles suggesting that business performance improve
Calculate probability that combinations of frequent expressions and a clue expression appear
Set of financial articles
Set of articles concerning business performance
Set of causal expressions
Assignment of polarity
Extraction of articles concerning business performance
Extraction of causal expressions
Examples
• ⟨positive⟩ Profit of software and service sector contributes.
• ⟨negative⟩ Slackness of textile sector affects the business performance.
Set of articles suggesting that business performance decline

Fig. 3 Outline of our method.

Table 1 Examples of words selected as features.

zyouhou syuusei: (upward adjustment)
kahou syusei: (downward adjustment)
tokubetu sonsitu: (extraordinary charge)
zousyuu zoueki: (profit increase)

and is calculated by the following Formula 6:

$$P(t_i, S_p) = \frac{Tf(t_i, S_p)}{\sum_{t \in Ts(S_p)} Tf(t, S_p)}. \tag{6}$$

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

The entropy $H(t_i, S_p)$ is based on probability $P(t_i, d)$ that word t_i appears in document $d \in S_p$, and is calculated by the following Formula 7:

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d), \tag{7}$$

$$P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d' \in S_p} tf(t_i, d')}. \tag{8}$$

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

Next, our method compares $W(t_i, S_p)$ with $W(t_i, S_n)$. If either score $W(t_i, S_p)$ is larger than $2W(t_i, S_n)$ or score $W(t_i, S_n)$ is larger than $2W(t_i, S_p)$, word t_i is selected as a feature for SVM. Some examples of words selected as features are shown in Table 1. Here, each element of feature vectors used for learning by SVM is the appearance probability of words selected as features contained in each article in training data. We use a linear kernel as the kernel of SVM.

4.2 Polarity Assignment to Causal Expressions by Using Combinations of Frequent Expressions and Clue Expressions

We explain about our method of assigning polarity to causal expressions. Here, we define a frequent expression as fp_i, a clue expression as cp, and a causal expression as $\boldsymbol{ce}$. Note that a causal expression contains a clue expression and at least one frequent expression.

$$\boldsymbol{ce} = (\langle fp_1, cp\rangle, \langle fp_2, cp\rangle, \ldots, \langle fp_n, cp\rangle). \tag{9}$$

Here, we define polarity of a causal expression as $c \in \{positive, negative\}$. Our method assigns polarity c to causal expression $\boldsymbol{ce}$ by calculating probability $P(c|\boldsymbol{ce})$ by the following Formula 10:

$$\hat{c} = \arg\max_c P(c|\boldsymbol{ce}) = \arg\max_c P(c)P(\boldsymbol{ce}|c) \tag{10}$$

However, probability $P(c)$ and $P(\boldsymbol{ce}|c)$ are not able to be calculated. Hence, our method estimates them by using the polarity of articles concerning business performance. The probability $P(\boldsymbol{ce}|c)$ is estimated by the following Formula 11:

$$P(\boldsymbol{ce}|c) \approx \prod_{i=1}^{n} P(\langle fp_i, cp\rangle | c_d), \tag{11}$$

where,

$c_d \in \{positive, negative\}$**:** the polarity of an article concerning business performance classified by our method.

$P(\langle fp_i, cp\rangle | c_d)$**:** the conditional probability that $\langle fp_i, cp\rangle$ contained in causal expressions appears in a set of articles concerning business performance classified to c_d.

Moreover, the probability $P(c)$ in Formula 10 is estimated by $P(c_d)$. Here, $P(c_d)$ is the probability that causal expressions appear in a set of articles concerning business performance classified to c_d.

Here, $P(c|\boldsymbol{ce})$ at $c = positive$ is defined as P_p and $P(c|\boldsymbol{ce})$ at $c = negative$ is defined as P_n, respectively. Our method assigns polarity "positive" in the case of $P_p > \beta P_n$ and assigns polarity "negative" in the case of $P_n > \beta P_p$. Otherwise, the polarity is not assigned.

5. Evaluation

5.1 Implementation

We implemented our method and evaluated it. We employ ChaSen[†] as a Japanese morphological analyzer, and

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

Table 2 Examples of causal expressions assigned polarity by our method.

causal expression	*tyuugoku muke no ekisyou kanren no densi buhin ga kaihuku suru*: (The amount of electronic parts related to liquid-crystal exported for China recover.)
frequent expression	*densi buhin*: (electronic parts), *ekisyou kanren*: (related to liquid-crystal)
clue expression	*ga kaihuku suru*: (recover)
polarity	positive
causal expression	*keitai zyouhou tanmatu ya densi debaisu ga husin datta*: (The amount of portable terminal and electronic device were down)
frequent expression	*densi debaisu*: (electronic device)
clue expression	*ga husin datta*: (were down)
polarity	negative

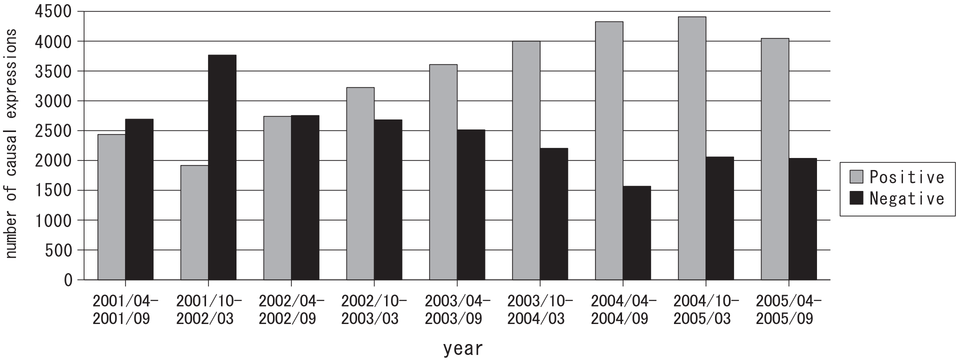

Fig. 4 The number of causal expressions assigned polarity by our method.

Table 3 Evaluation results.

β	$P_{pos}(\%)$	$R_{pos}(\%)$	F_{pos}	$P_{neg}(\%)$	$R_{neg}(\%)$	F_{neg}
1	68.7	51.0	58.6	73.6	63.0	67.9
1.5	74.4	50.4	60.0	76.8	61.5	68.3
2	75.3	47.9	58.6	77.0	58.5	66.5
2.5	75.4	46.5	57.6	76.8	55.5	64.4
3	76.1	45.1	56.7	78.1	53.1	63.2

CaboCha[†] as a Japanese parser and SVM^{light}[††] as an implementation of SVM. As training data, we manually extracted 2,920 articles concerning business performance from Nikkei newspapers published in 2000 and manually classified them into positive and negative ones. Causal expressions are extracted from Nikkei newspapers published from 1990 to 2005 (except 2000) by our previous method [1] and our method assigned polarity, "positive" or "negative", to them[†††]. Note that parameter α used for determining the threshold value in Formula 2 was 0.3. In this case, our previous method attained 81.8% precision and 67.9% recall, respectively.

Some examples of causal expressions assigned polarity by our method are shown in Table 2. Moreover, the number of causal expressions assigned polarity positive by our method and the number of causal expressions assigned polarity negative by our method are shown in Fig. 4. Note that business condition of Japan was marked bottom in 2001 and has gradually recovered since 2002.

5.2 Evaluation Results

As a correct data set for evaluation, we manually extracted 623 causal expressions from 138 articles concerning business performance and assigned polarity to them. After that, our method extracted causal expressions and assigned polarity to them by our method from the same 138 articles as test data and we calculated precision and recall. Table 3 shows the results. Here, values P_{pos} (P_{neg}) and R_{pos} (R_{neg}) are precision and recall of causal expressions to which our method assigns polarity positive (negative), respectively. Value F_{pos} (F_{neg}) is F measure of P_{pos} (P_{neg}) and R_{pos} (R_{neg}). Parameter β is a threshold value used for determining polarity in Sect. 4.2. Precision P_{pos} and recall R_{pos} is calculated by the following Formula 12:

$$P_{pos} = \frac{|A_{pos} \cap C_{pos}|}{|A_{pos}|}, \quad R_{pos} = \frac{|A_{pos} \cap C_{pos}|}{|C_{pos}|} \tag{12}$$

Here, A_{pos} is the set of causal expressions to which our method assigns polarity positive and C_{pos} is the set of causal expressions assigned polarity positive in a correct data set. Moreover, we evaluate our method when the number of

[†] http://chasen.org/˜taku/software/cabocha/

[††] http://svmlight.joachims.org

[†††] Although inappropriate causal expressions may be contained, they are not eliminated by hand.

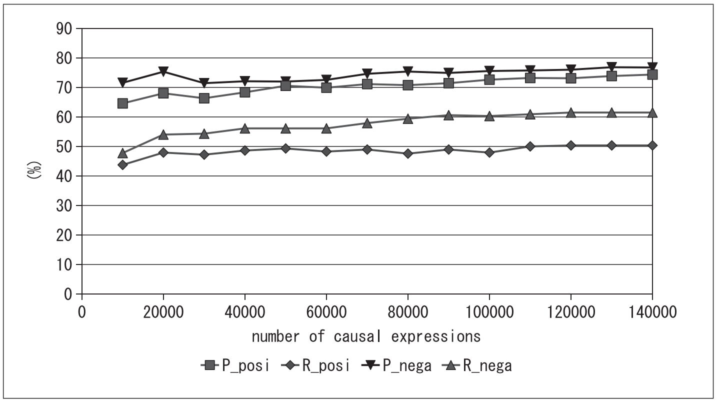

Fig. 5 Evaluation results when the number of causal expressions is changed.

Table 4 Evaluation results comparing our method with other methods.

	$P_{pos}(\%)$	$R_{pos}(\%)$	F_{pos}	$P_{neg}(\%)$	$R_{neg}(\%)$	F_{neg}
Our method	74.4	50.4	60.0	76.8	61.5	68.3
Average of $SO(\langle fp_i, cp\rangle)$	67.1	34.7	45.8	58.4	59.1	58.8
Average of $PV_{PMI}(\langle fp_i, cp\rangle)$	77.5	39.9	52.7	77.6	50.5	61.2
baseline (article)	64.0	43.8	52.0	62.1	53.7	57.6
baseline (dictionary)	100.0	10.4	18.9	100.0	1.5	2.9
baseline (score)	65.3	43.8	52.4	63.0	53.4	57.8

causal expressions for calculating probability that combinations of frequent expressions and a clue expression appear in a set of positive articles or that of negative ones is changed. Figure 5 shows the results.

5.3 Comparison of Our Method with Other Methods

We compare our method with a method based on the average of $SO(\langle fp_i, cp\rangle)$, a method based on the average of $PV_{PMI}(\langle fp_i, cp\rangle)$, and baseline methods (article, dictionary, score). Table 4 shows the results. For the purpose of reference, we also shows precision and recall of our method in Table 4. Here, we set threshold value β as 1.5. These comparison methods are explained in the following subsection, respectively.

5.3.1 Comparison with a Method Based on the Average of $SO(\langle fp_i, cp\rangle)$

We compare our method with a method based on Semantic Orientation (*SO*) that is a criterion proposed by Turney [4]. The Semantic Orientation (*SO*) is calculated by the following Formula 13.

$$SO(ce) = \log_2 \frac{hits(ce\ NEAR\ \text{``}koutyou\text{''})hits(\text{``}husin\text{''})}{hits(ce\ NEAR\ \text{``}husin\text{''})hits(\text{``}koutyou\text{''})} \quad (13)$$

Note that we used "*koutyou* (good)" and "*husin* (down)", respectively, instead of "excellent" and "poor" used by Turney for calculating *SO*. Actually, the method proposed by Turney is a method for classifying reviews as *recommended* (thumbs up) or *not recommended* (thumbs down) by using average *SO* of phrases contained in the reviews. Here, we set a comparison method that assigns polarity to causal expressions by using average *SO* of combinations of a frequent expression and a clue expression contained in the causal expressions. $SO(\langle fp_i, cp\rangle)$ of a combination of frequent expression fp_i and clue expression cp is calculated by the following Formula 14.

$$SO(\langle fp_i, cp\rangle) = \log_2 \frac{hits(\langle fp_i, cp\rangle\ NEAR\ \text{``}koutyou\text{''})hits(\text{``}husin\text{''})}{hits(\langle fp_i, cp\rangle\ NEAR\ \text{``}husin\text{''})hits(\text{``}koutyou\text{''})} \quad (14)$$

Here, $hits(\langle fp_i, cp\rangle\ NEAR\ \text{``}koutyou\text{''})$ is a number of articles containing both a combination of frequent expression fp_i and clue expression cp and "*koutyou* (good)" within 22 words†.

5.3.2 Comparison with a Method Based on the Average of $PV_{PMI}(\langle fp_i, cp\rangle)$

We compare our method with a method based on the average of $PV_{PMI}(\langle fp_i, cp\rangle)$. Kaji et al. proposed a method for acquiring polar phrases (adjectives and "noun + postpositional particles + adjective") by using frequency in polar sentence corpus and *PMI* (Pointwise Mutual Information) based polarity value [11]. The PV_{PMI} based polarity value is calculated by the following Formula 15.

$$PV_{PMI}(ce) = PMI(ce, pos) - PMI(ce, neg) = \log_2 \frac{P(ce|pos)}{P(ce|neg)} \quad (15)$$

†22 is the average of the number of words that compose a causal expressions.

Here, $P(ce|pos)$ is ce's probability in positive sentences. However, since the polar sentences are extracted by using lexicon-syntactic patterns and a manually created cue words list, they are not applicable to our task. Hence, we compare our method with a method based on the average of $PV_{PMI}(\langle fp_i, cp\rangle)$ as the following Formula 16.

$$PV_{PMI}(\langle fp_i, cp\rangle) = \log_2 \frac{P(\langle fp_i, cp\rangle|pos)}{P(\langle fp_i, cp\rangle|neg)} \qquad (16)$$

Here, $P(\langle fp_i, cp\rangle|pos)$ $(P(\langle fp_i, cp\rangle|neg))$ is $\langle fp_i, cp\rangle$'s probability in positive (negative) sentences. We set $P(\langle fp_i, cp\rangle|pos)$ $(P(\langle fp_i, cp\rangle|neg))$ as $\langle fp_i, cp\rangle$'s probability in positive (negative) articles concerning business performance in this evaluation. Moreover, we set threshold value θ for deciding polarity as 1.0.

5.3.3 Comparison with Baseline Methods

As a baseline method (article), we evaluate the method that assigns polarity corresponding to polarity of an article containing a causal expression to the causal expression. (For example, if an article containing a causal expression "*zidousya no uriage ga koutyou*: (Sales of cars are good)" is classified into negative one, polarity of the causal expression is also negative.)

As a baseline method (dictionary), we evaluate the method based on the annotation of dictionary. Actually, the method assigns polarity positive to causal expressions containing a clue expression "*ga koutyou*: (be good)" and assigns polarity negative to causal expressions containing a clue expression "*ga husin*: (be down)".

We consider that replacing a causal expression with fp_i and cp mainly contributes to the improvement of our method. To confirm it, we compared our method with a baseline method (score) based on the following score.

$$score = \log_2 \frac{P(pos)P(ce|pos)}{P(neg)P(ce|neg)} \qquad (17)$$

Here, $P(ce|pos)$ is ce's probability in positive articles concerning business performance and $P(pos)$ is the probability that causal expressions appear in a set of positive articles concerning business performance.

6. Discussion

Table 4 shows that our method attained 74.4% precision of assigning polarity positive and 76.8% precision of assigning polarity negative to causal expressions extracted by our previous method. Note that if our method assigns polarity to inappropriate causal expressions extracted by our previous method (the precision of extracting causal expressions was 81.8%.), they are recognized as incorrect[†]. Our method attained 88.4% precision of assigning polarity positive and 92.5% precision of assigning polarity negative if it assigns polarity to appropriate causal expressions selected manually in order to evaluate only our method. Hence, we consider that our method achieved good performance. For example, our method was able to assign polarity positive to a causal expression that contains a frequent expression "*uriage*: (sales)" and a clue expression "*ga zouka sita*: (are increasing)". Moreover, our method was able to assign polarity negative to a causal expression that contains a frequent expression "*shoukyaku hutan*: (extinguishment responsibility)" and a clue expression "*ga zouka sita*: (are increasing)". In general, the polarity of a causal expression containing a clue expression "*ga koutyou*: (are good)" is positive and the polarity of a causal expression containing a clue expression "*ga husin*: (are down)" is negative. However, the polarity is not able to be determined only by clue expressions e.g., "*ga zouka*: (are increasing)". Our method classifies articles concerning business performance into positive and negative ones, and assigns polarity to a causal expression by using probability that combinations of frequent expressions and a clue expression appear in the set of positive articles or in that of negative ones. Hence, our method was able to assign appropriate polarity even if the polarity to be assigned is changed by the combination of frequent expressions and clue expressions.

In our method, the precision of classifying articles concerning business performance influences the precision of assigning polarity to causal expressions. We evaluated the method of classifying articles concerning business performance and it attained 89.8% precision[††, †††]. We consider that the method of classifying articles concerning business performance achieved good performance since our method is able to extract features efficient for classifying articles concerning business performance, e.g., "*zyouhou syuusei*: (upward adjustment)", "*tokubetu sonsitu*: (extraordinary charge)". Moreover, it attained 88.4% precision even if content words effective for classifying articles were not selected as features (i.e., all words contained in articles are selected as features.). As mentioned above, we consider that it is easy to classify articles concerning business performance into positive and negative ones. In contrast, since a causal expression consists of some words (the average of the number of words that compose a causal expressions is 21.5.), the number of content words effective for assigning polarity is low. Hence, we consider that it is more difficult to assign polarity to causal expressions than to classify articles concerning business performance into positive and negative ones. Our method uses frequent expressions and clue expressions acquired to extract causal expressions as information to assign them polarity. We consider that our method achieved good performance by using the frequent expressions and the clue expressions.

[†] In this evaluation, we do not exclude inappropriate causal expressions manually.

[††] As a correct data set for calculating the precision of classifying articles, we manually classified 550 articles concerning business performance into positive and negative ones.

[†††] We compared a method without any feature selection (i.e., all words contained in articles are selected as features.) with our method. As a result, the method without any feature selection attained 88.4% precision.

The recall value of assigning polarity positive and that of assigning polarity negative were 50.4% and 61.5%, respectively. The reason why the recall value is low is that the recall value of extracting causal expressions extracted by our previous method is 67.9%. Moreover, our method assigns polarity positive in the case of $P_p > \beta P_n$ and assigns polarity negative in the case of $P_n > \beta P_p$, otherwise, the polarity was not assigned. If our method does not execute this processing (i.e., $\beta = 1$), the precision value and the recall value of assigning polarity positive were 68.7% and 51.0%, and the precision value and the recall value of assigning polarity negative were 73.6% and 63.0%, respectively. Hence, although the precision was improved, the recall was decreased by this processing.

Table 4 shows that our method outperformed the method based on the average of $SO(\langle fp_i, cp\rangle)$. $SO(\langle fp_i, cp\rangle)$ is calculated by using frequency of co-occurrence of $\langle fp_i, cp\rangle$ and "*koutyou* (good)", and "*husin* (down)", respectively. However, since there are a lot of clue expressions effective for extracting causal expressions, a lot of causal expressions that do not co-occur with "*koutyou* (good)" or "*husin* (down)" exist. Hence, $SO(\langle fp_i, cp\rangle)$ was not able to be calculated appropriately. Hence, we consider that our method outperformed the method based on the average of $SO(\langle fp_i, cp\rangle)$.

Table 4 shows that, compared with our method, the precision of the average of PV_{PMI} is increased, but the recall is decreased. The reason why the recall of the average of PV_{PMI} is low is that the case where the polarity is not assigned has increased since the difference between $P(\langle fp_i, cp\rangle|pos)$ and $P(\langle fp_i, cp\rangle|neg)$ is small. Since a causal expression consists of a number of words (the average of the number of words that compose a causal expressions is 21.5.), the same causal expression does not appear in the set of articles concerning business performance. Here, our method solves the problem that the same causal expression does not appear by replacing a causal expression that consists of a number of words with some "frequent expressions" and a "clue expression". Moreover, our method assigns polarity by using probability that combinations of frequent expressions and a clue expression appear in a set of positive articles or that of negative ones. These processes are features of our method. However, all features of our method are introduced into the method based on the average of $PV_{PMI}(\langle fp_i, cp\rangle)$. Hence, we consider that it also attained high precision.

Table 4 shows that the precision of assigning polarity positive and the precision of assigning polarity negative of the baseline method (article) were 64.0% and 62.1%, respectively. The reason why precision of the baseline method is low is that causal expressions that should assign polarity negative are frequently contained in articles classified into positive.

Table 4 shows that the precision of baseline (dictionary) is high, but the recall is very low. The reason why the recall of baseline (dictionary) is very low is that the method based on the annotation of dictionary is not able to assign polarity to causal expressions since the number of clue expressions, i.e., 2, is too small[†].

Table 4 shows that our method outperformed the baseline method (score). Since a causal expression consists of a number of words, the same causal expression does not appear in the set of articles concerning business performance. Hence, $P(ce|pos)$ and $P(ce|neg)$ in the score was not able to be calculated appropriately. In contrast, our method solves the problem that the same causal expression does not appear by replacing a causal expression that consists of a number of words with some "frequent expressions" and a "clue expression". Hence, we consider that our method outperforms the baseline method (score).

7. Conclusion

In this paper, we proposed a method of assigning polarity (positive or negative) to causal expressions extracted from articles concerning business performance of companies. Our method assigned polarity to them by using statistical information. First, our method classified articles concerning business performance into positive and negative ones. Next, our method assigned polarity to a causal expression by using probability that combinations of frequent expressions and clue expressions appear in a set of positive articles or that of negative ones. We evaluated our method and it attained 74.4% precision and 50.4% recall of assigning polarity positive, and 76.8% precision and 61.5% recall of assigning polarity negative.

As a future direction, we consider a task to assign an importance score to a causal expression by using the polarity assigned by our method. That is, a causal expression related to its core business (e.g., sales of cars are good) is assigned a large score and a causal expression not related to its core business (e.g., profit from sales of stocks) is assigned a small score. By assigning the importance score, we are able to provide the causal expression as more useful information for supporting decision on investment to private investors. Moreover, the polarity assigned by our method is used for identifying whether the importance score is positive or negative.

Acknowledgment

This work was supported in part by Grant-in-Aid for Scientific Research (B) (20300058) and Grant-in-Aid for Young Scientists (B) (21700158) from Japan Society for the Promotion of Science (JSPS), Research for Promoting Technological Seeds from Japan Science and Technology Agency (JST), The Telecommunications Advancement Foundation and Kayamori Foundation of Informational Science Advancement.

[†] Note that the number of clue expressions and frequent expressions acquired from Nikkei newspapers published from 1990 to 2005 (except 2000) reaches 1,026 and 1,071, respectively.

References

[1] H. Sakai and S. Masuyama, “Cause information extraction from financial articles concerning business performance,” IEICE Trans. Inf. & Syst., vol.E91-D, no.4, pp.959–968, April 2008.

[2] H. Takamura, T. Inui, and M. Okumura, “Latent variable models for semantic orientations of phrases,” Proc. 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006), pp.201–208, 2006.

[3] H. Sakaji, H. Sakai, and S. Masuyama, “Automatic extraction of basis expressions that indicate economic trends,” Proc. PAKDD2008, pp.977–984, 2008.

[4] P.D. Turney, “Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews,” Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL2002), pp.417–424, 2002.

[5] R.E. Schapire and Y. Singer, “Boostexter: A boosting based system for text categorization,” Machine Learning, vol.39, no.2/3, pp.135–168, 2000.

[6] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” Proc. Joint Conference on Human Language Technology/Conference on Empirical Methods in Nutural Language Processing (HLT/EMNLP ’05), pp.347–354, 2005.

[7] F. Smadja, “Retrieving collocations from text: Xtract,” Computational Linguistics, vol.19, no.1, pp.143–177, 1993.

[8] F. Baron and G. Hirst, “Collocations as cues to semantic orientation,” AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI-EAAT 2004), 2004.

[9] H. Takamura, T. Inui, and M. Okumura, “Extracting semantic orientations of phrases from dictionary,” The Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT2007), pp.292–299, 2007.

[10] F.Y. Wu, “The potts model,” Reviews of Modern Physics, vol.54, no.1, pp.235–268, 1982.

[11] N. Kaji and M. Kitsuregawa, “Building lexicon for sentiment analysis from massive collection of html documents,” Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp.1075–1083, 2007.

[12] 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.

[13] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, “Mining of concurrent text and time series,” Proc. KDD 2000 Conference Text Mining Workshop, 2001.

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.