Automatic Extraction of Basis Expressions that Indicate Economic Trends

Hiroki Sakaji[1], Hiroyuki Sakai[1], and Shigeru Masuyama[1]

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

Abstract. This paper proposes a method to automatically extract basis expressions that indicate economic trends from newspaper articles by using a statistical method. We also propose a method to classify them into positive expressions that indicate upbeat, and negative expressions that indicate downturn in economy, respectively. It is important for companies, governments and investors to predict economic trends in order to forecast revenue, sales of products, prices of commodities and stock prices. We considered that basis expressions are useful for the companies, governments and investors to forecast economic trends. We extracted basis expressions, and classified them into positive expressions or negative expressions as information to forecast economic trends. Our method used a bootstrap method that was minimally a supervised algorithm for extracting basis expressions. Moreover, our method classified basis expressions into positive expressions or negative ones without dictionaries.

1 Introduction

It is important for companies, governments and investors to predict the economic trends in order to forecast revenue, sales of products, prices of commodities and stock prices. The diffusion index[1] is one of indices concerning economic trends, and is computed every three months, and provides economic trends during prior period. However, it is difficult to forecast the business performance accurately by using diffusion indices, as it can not indicate current economic trends.

These indices are computed using numeric data. However, some qualitative language data that reflect economic trends may not be quantified straightforwardly. For example, an opinion "*Economy seems to recover*" in a newspaper article is hard to be quantified, as "*Economy seems to recover*" is a sense of the writer.

Nakajima et al.[1] proposed a method for extracting articles concerning economic trends from newspaper articles and classifying them into positive articles

[1] The diffusion index is a summary measures designed to facilitate the analysis and forecast of business cycles by combining the behavior of a group of economic indicators that represent widely differing activities of the economy, such as production and employment, and that correspond closely to turning points.
http://www.esri.cao.go.jp/en/stat/di/di2e.html

that indicate upbeat in economy and negative ones that indicate downturn in economy. However, Nakajima's method can not classify articles having two different opinions. For example, an article that indicates economy in Aichi prefecture is upbeat while that in Gifu prefecture is downturn, includes two different opinions about different areas, and can not be treated by Nakajima's method.

We propose a method to extract basis expressions that indicate economic trends from newspaper articles concerning economic trends and to classify basis expressions into positive or negative expressions. We considered that opinions concerning economic trends can be extracted by using basis expressions, which enable us to distinguish two different types of opinions in the same articles. Our method used a bootstrap method that was minimally supervised algorithm for extracting basis expressions. Moreover, our method classified basis expressions into positive expressions or negative ones without dictionaries.

2 Related Work

As related work for extracting phrases that have a particular meaning, 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[5]. 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. In contrast, our method uses statistical information and only one initial clue phrase as an initial input. The domain-specific dictionaries, predetermined patterns, complex rules made by hand are not needed.

Wilson et al. proposed a method for determining whether an expression is neutral or polar[6]. In their research, the expressions are extracted manually and the method needs dictionaries. In contrast, our method automatically extracts expressions and does not need dictionaries.

Sakai et al. proposed a method for extracting cause information from Japanese financial articles concerning business performance[3]. Their work is probably most closely related to ours. However, our method extracts basis expression concerning not performance of each company but economic trends. Moreover, our method also classifies basis expressions into positive and negative ones.

3 Extraction of Basis Expressions

As a preprocessing, our method extracts articles concerning economic trends from newspaper corpus by using Support Vector Machine(SVM)[4]. We applied a method proposed by sakai et al.[2] for extracting them. As a result, 10,027 newspaper articles concerning economic trends were extracted from Nikkei newspapers published from 1990 to 2005.

Here, a basis 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 basis expressions by using clue phrases, i.e. phrases frequently modified by basis expressions. For example, a basis expression frequently modifies clue

phrase "の影響 (*no eikyou*: influenced by)" in Japanese. Our method extracts an expression that consists of a clue phrase and a phrase that modifies it as a basis expression. Hence, if many clue phrases effective for extracting basis expressions are acquirable, basis expressions are extracted automatically. However, it is hard to acquire sufficient clue phrases effective for extracting basis expressions manually. Hence, our method also acquires such clue phrases automatically from a set of articles concerning economic trends.

Our method for extracting basis expressions is as follows.

Step 1: Input an initial clue phrase "の影響 (*no eikyou*: influenced by)" and acquire phrases that modify them.
Step 2: Extract phrases appearing frequently in a set of the phrases acquired in Step 1 (e.g. 世界経済 (*sekai keizai*: world economy)). In this paper, such a phrase extracted in Step 2 is defined as a "frequent phrase".
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 they are executed predetermined times or neither new clue phrases nor new frequent phrases are extracted.
Step 6: Extract basis expressions by using extracted frequent phrases and acquired clue phrases.

3.1 Extraction of Frequent Phrases

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

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

Score $S_f(e, c)$ is calculated by the following Formula 1:

$$S_f(e, c) = -f_e(e, c) f_p(e) \log_2 P(e, c), \tag{1}$$

where $P(e, c)$ is the probability that frequent phrase candidate e containing c appears in the set of articles concerning economic trends. $f_e(e, c)$ is the number of frequent phrase candidate e's containing c in the set of articles concerning economic trends. $f_p(e)$ is the number of *bunsetu*'s that compose e. $P(e, c)$ is calculated by the following Formula 2.

$$P(e, c) = \frac{f_e(e, c)}{Ne(c)}, \tag{2}$$

where $Ne(c)$ is the total number of frequent phrase candidates containing c in the set of articles concerning economic trends.

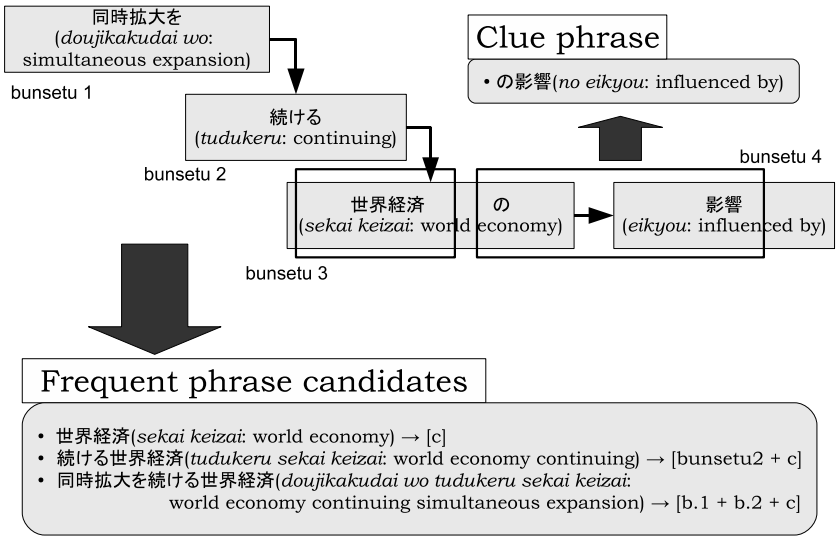

Fig. 1. Examples of frequent phrase candidates

3.2 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 $P(e, s)$ and selects frequent phrases assigned entropy $H(e)$ larger than a threshold value calculated by Formula 5. $P(e, s)$ is the probability that frequent phrase e modifies clue phrase s. 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 an appropriate frequent phrase. Entropy $H(e)$ is calculated by the following Formula 3.

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

where

$$P(e, s) = \frac{f(e, s)}{\sum_{s' \in S(e)} f(e, s')}. \tag{4}$$

Here, $S(e)$ is the set of clue phrases modified by frequent phrase e. $f(e, s)$ is the number of frequent phrase e's that modifies clue phrase s in the set of articles concerning economic trends. 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 α is a constant ($0 < \alpha < 1$).

3.3 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 Formula 7.

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

$$H(s) = -\sum_{e \in E(s)} P(s,e) \log_2 P(s,e), \qquad P(s,e) = \frac{f(s,e)}{\sum_{e' \in E(s)} f(s,e')}. \tag{6}$$

Here, $E(s)$ is the set of frequent phrases that modify clue phrase s, and $f(s, e)$ is the number of clue phrase s's modified by frequent phrase e in the set of articles concerning economic trends. The threshold value is calculated by the following Formula 7.

$$T_s = \alpha \log_2 |N_e|. \tag{7}$$

Here, N_e is the set of frequent phrases used for extracting clue phrases. α is the same constant that in Formula 5.

3.4 Extraction of Basis Expressions by Using Frequent Phrases and Clue Phrases

Finally, our method extracts basis expressions by using frequent phrases and clue phrases. A basis expression consists of a phrase that modifies the clue phrase. Moreover, the phrase that modifies the clue phrase contains some frequent phrases. For example, "輸出の減少を背景に (*yusyutu no gennsyou wo haikei ni*: under decreasing export)" is a basis expression since phrase "輸出の減少 (*yusyutu no gennsyou*: decreasing export)" modifies clue phrase "を背景に (*wo haikei ni*: under)" and the phrase contains frequent phrase "減少 (*gensyou*: decreasing)".

4 Classification of Basis Expressions

Our method classifies extracted basis expressions into positive expressions and negative expressions. However, extracted basis expressions contain some of inappropriate basis expressions. As a result, our method extracted basis expressions into positive expressions, negative expressions and other expressions. Other expressions are extracted basis expressions that are neither positive nor negative expressions.

For example, "同時拡大を続ける世界経済 (*doujikakudai wo tudukeru sekaikeizai*: world economy continuing simultaneous expansion)" is a positive expression. Thus positive expressions indicate that Japanese economy is upbeat. For example, "設備投資や個人消費の鈍化 (*setubitoushi ya kojinsyouhi no donka*: slowdown of business investment and personal consumption)" is a negative expression. Thus negative expressions indicate that Japanese economy is downturn. For example, "調査対象変更 (*tyousataisyouhenkou*: change of objective for survey)" and "景気の伸び悩み (*keiki no nobinayami*: stagnation of economy)" are other expressions. We define expressions that cite Japanese economy are inappropriate as basis expressions, because our goal is extraction of basis expressions.

We develop two classifiers by using one-versus-rest method and Support Vector Machine(SVM)[4]. The one classifies extracted basis expressions into positive expressions and the others. Here, a positive expression is defined as a *correct expression* and the other is defined as an *incorrect expression*. The other one classifies extracted basis expressions into negative expressions and the others. Here, negative expression is defined as a *correct expression* and the other is defined as an *incorrect expression*. The classifiers use *character N-gram* and *word N-gram* as features.

5 Evaluation

In this section, we evaluate our method. Our method extracted basis expressions from 10,027 newspaper articles concerning economic trends and classify them into positive and negative.

First, we evaluated our method for extracting basis expressions. We employ CaboCha[2] as a Japanese parser. We manually extracted 75 basis expressions from 100 articles concerning economic trends performance as a correct data set. Moreover, we extracted basis expressions by our method from the same 100 articles and calculated precision and recall. Here, a basis expression extracted by our method is correct if it contains a basis expression extracted as the correct data set. The precision, recall and F-measure[3] calculated by the following formulas.

$$Precision = \frac{|Sb \cap Ab|}{|Sb \cap Nb|},\ Recall = \frac{|Sb \cap Ab|}{|Ab|},$$

where Sb is the set of basis expressions extracted by our method from 100 articles concerning economic trends. Ab is the set of basis expressions contained in the correct data set. Nb is the set of expressions modifying clue phrases in the 100 articles concerning economic trends. The results are shown in Tabel 1.

Next, we evaluated our method for classifying basis expressions. We employ ChaSen[4] as a Japanese morphological analyzer, and SVM^{light}[5] as an implementation of SVM. We extracted 1620 basis expressions by our method with α

[2] http://chasen.org/~taku/software/cabocha/
[3] $F - measure = (2 \times Precision \times Recall)/(Precision + Recall)$
[4] http://chasen.aist-nara.ac.jp/hiki/ChaSen/
[5] http://svmlight.joachims.org

Table 1. Precision, Recall and F-measure of basis expressions extraction

α	Precision	Recall	F-measure	num. of basis expression
0.9	1.000	0.160	0.276	650
0.6	0.714	0.333	0.455	1620
0.5	0.042	0.573	0.078	49293

Table 2. Precision, recall and F-measure of basis expression classification with *character N-gram feature*

	num. of frequent features	Precision	Recall	F-measure
Positive	9021	0.800	0.615	0.695
Negative	9021	0.843	0.855	0.849

0.6 and iteration count 3. 1620 basis expressions were manually annotated with "positive", "negative" or "others". The annotated basis expressions were divided into two sets. The first (1120 expressions) were a training data, used for feature selection and modeling. We used the second set (500 expressions) as a test data set. We calculated precision, recall and F-measure from the test data set. The precision and recall are calculated by the following formulas.

$$Precision = \frac{|E \cap C|}{|E|},\ Recall = \frac{|E \cap C|}{|C|}.$$

Here, E is the set of basis expressions annotated with *correct expressions* in the test data set. C is the set of *correct expressions* contained in the test data set. The results are shown in Tables 2 and 3.

6 Discussions

In Table 1, precision rises from α 0.5 to 0.6, while recall drops. When low α value is assigned, inappropriate clue phrases and frequent phrases were found in a set of extracted clue phrases and extracted frequent phrases. Furthermore, new inappropriate ones are extracted by extracted inappropriate phrases. As a result, our method acquires many inappropriate ones. This happens when α is between 0.5 and 0.6.

In Tables 2 and 3, focusing on recall, it is interesting to note that *negative* classifiers perform better than *positive* ones. This is due to the fact that 1620 basis expressions contain negative expressions far more than positive ones. We consider that cause for few positive expressions is due to recession in Japan during the period of the corpus.

In Tables 2 and 3, focusing on precision, *character N-gram feature* classifier is the highest of all. This is due to the fact that characters play a key role in classifying expressions in Japanese, because Chinese characters that are one of the Japanese characters have meanings.

Table 3. Precision, recall and F-measure of basis expression classification with *word N-gram feature*

	num. of frequent features	Precision	Recall	F-measure
Positive	6450	0.769	0.545	0.638
Negative	6450	0.833	0.867	0.845

7 Conclusion

We proposed a method for extracting basis expressions that indicate economic trends from Japanese newspaper articles concerning economic trends. First, our method extracts basis expressions from them by using statistical information and initial clue phrases. Next, our method classifies basis expressions into *positive expressions*, *negative expressions* and *other expressions*. This method can also be applied to other tasks such as extracting reputations for specific items.

Acknowledgment

This work was supported in part by Global COE Program "Frontiers of Intelligent Sensing ", MEXT, Japan.

References

1. Takumi Nakajima, Hiroyuki Sakai and Shigeru Masuyama, "A classification method based on the view of the author of each newspaper article on economics", *IPSJ SIG Notes*, Vol.2003, No.51(20030522), pp.175-180, 2003(in Japanese).
2. Hiroyuki Sakai, Shouji Umemura and Shigeru 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(in Japanese).
3. Hiroyuki Sakai and Shigeru Masuyama, "Extraction of Cause Information from Newspaper Articles Concerning Business Performance", Proc. of the 4th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI2007), pp.205-212 , 2007.
4. V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1999.
5. Hiroshi Kanayama, Tetsuya Nasukawa and Hideo Watanabe, "Deeper sentiment analysis using machine translation technology", Proceedings of the 20th COLING, pp.494-500, 2004.
6. Theresa Wilson, Janyce Wiebe and Paul Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", In *Proceedings of HLT/EMNLP-05*, pp.347-354, 2005.