

A “Catchy” Copy and Concept Evaluation System Using a Natural Language Processing Approach

Sadahiro IKEDA and Shigeo KANEDA

The Graduate School of Policy and Management, Doshisha University
Karasuma-Imadegawa, Kyoto-shi, 602-8580, Japan

ABSTRACT

In advertising, marketing research, CF (Commercial Film) creation, and drama creation, it is very important to be able to predict the words or concepts that will come into vogue. Initially, a creator comes up with many candidates for future vogue words or concepts. Next, the creator selects the best word or concept for the most “catchy” copy. This selection process is very hard because it is essentially a sensory and intuitive job. Thus, the creator often cannot explain why he/she thinks a particular candidate word is the best one. To resolve this problem, this paper demonstrates a new approach using natural language processing techniques and the “Contemporary Word Dictionary” (Jiyukokumin-sha, Tokyo) to support creators in selecting the best word or concept from several candidates. This system is based on the assumption that a vogue word or concept is a reflection of the social, economic, or political situation. Using this system, the “Most Popular Vogue Words” selected in the year 2000 by the publishing company Jiyukokumin-sha were evaluated by comparison with the former “Most Popular Vogue Words” selected in the years 1997–1999. The experimental results obtained were also tested statistically. The statistical test showed that the result was expressive. This means that the proposed approach is both practical and effective.

Keywords: Support System, Word in Vogue, Natural Language Processing, Contemporary Word Dictionary, Statistical Test.

1. INTRODUCTION

In advertising, marketing research, CF (Commercial Film) creation, and drama creation, it is very important to be able to predict the words or concepts that will come into vogue. Initially, a creator comes up with many candidates for future vogue words or concepts. Next, the creator selects the best word or concept for the most “catchy” copy. This selection process is very hard because it is essentially a sensory

and intuitive job. Thus, the creator often cannot explain why he/she thinks a particular candidate word is the best one.

To resolve this problem, this paper demonstrates a new approach using natural language processing techniques and the “Contemporary Word Dictionary” (Jiyukokumin-sha, Tokyo) to support creators in selecting the best word or concept from several candidates.

This system is based on the assumption that a vogue word or concept is a reflection of the social, economic, or political background. Using this system, the “Most Popular Vogue Words” selected in the year 2000 by the publishing company Jiyukokumin-sha were evaluated by comparison with the former “Most Popular Vogue Words” selected in the years 1997–1999. The experimental results obtained were also tested statistically. The statistical test showed that the result was expressive. This means that proposed approach is both practical and effective.

The following Section 2 analyzes the relationship between the social situation and a word or concept in vogue. Section 3 presents a new system to predict a future vogue word. Section 4 demonstrates experimental results. Section 5 shows a financial engineering approach. Section 6 concludes this paper.

2. VOGUE WORD OR CONCEPT AND SOCIAL BACKGROUND

“Shomuni” was a very popular Japanese TV drama, which was broadcast in 1998. The word “Shomuni” means “second group of the General Affairs Section.” Usually, jobs performed by members of a general affairs section in a company are simple. Use of the word “second” means that the section is not major one. The group leader of this group is quite a negative man whose major jobs at the office are having a midday nap and keeping a big cat.

In this TV drama, all the group members, except the group leader, are young women whose uniforms are short miniskirts. The main jobs of the group are replacing fluorescent lamps and toilet paper, cleaning toilets, and planning recreation trips for employees.

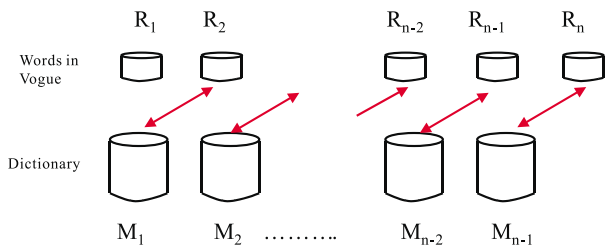


Figure 1: System Block Diagram

The stage settings of this show are, of course, quite different from the real general affairs sections of Japanese companies and show that the women are considered to be unworthy and hostile members in the company.

The “Shomuni” young women are very active in company matters. They often put forward just arguments and solve many big problems for the company. They change many “impossible situations” into possible ones. For example, they remonstrate with one of their company’s important customers about his selfish demands. They do not get on well with a young and beautiful secretary of the company president. In fact, they hate her. But they battle with the misunderstanding that she is a woman of loose morals. The following aspects of the social situation in Japan can be considered as reasons why the drama “Shomuni” was such a big hit.

- Most employees are afraid that they will lose their jobs because of the long-term depression of Japanese economy. On the other hand, they cannot express their dissatisfaction with a management’s lack of policy.
- The speech of the “Shomuni” women is very rough. But, they are sympathetic with weak employees. This is a typical image of the old Japanese male culture.
- Japan has recently experienced a greater participation by women in public affairs, for example, with the election of female city and prefectural governors. Young Japanese female employees are no longer satisfied with the role of “office lady,” serving tea and making photocopies.

We could easily put forward many more reasons for the popularity of this show. It can easily be shown that a vogue word or concept depends upon the social, economic, or political situation. However, there is no “proof” concerning the relationship between the social situation and the vogue word or concept. This “proof” is essentially sensory and intuitive.

This results in a big problem for the creator. For example, creators or planners of CF (Commercial Film) come up with several candidates for concepts or catchy

copy for a client company. The creator selects the best one from the candidates. However, he/she cannot explain clearly why the one they have chosen is the best. Thus, the client company’s manager finds it hard to accept their proposal.

If the relationship between the social situation and the word in vogue can be clarified quantitatively, the creator can argue that the created catchy copy will be effective. Quantitative evaluation is a powerful tool for the CF competition at the client’s office because the client company’s manager can understand why a particular candidate has been selected.

The major aim of this paper is to find a new approach to get a quantitative analysis of the relationship between the social situation and the candidate catchy copies created by creators. If the creators employ the proposed system, they can explain why a particular candidate word will come into vogue in the future. Of course, we cannot predict the future. Our quantitative analysis system is a kind of support tool. A good result in the proposed system shows that there is a high probability that the candidate word come into vogue.

3. EVALUATION SYSTEM FOR WORDS AND CONCEPTS IN VOGUE

3.1 OUTLINE OF THE PROPOSED SYSTEM

This system is based on the assumption that vogue words or concepts depend upon the current social, economic, or political situation. This social situation is expressed by “words” in this system. The system calculates the distance between the created catchy copy candidates and the “words” of the social situation. The system employs a contemporary word dictionary for the “words” of the social situation.

Contemporary Word Dictionary: The system employs the “Contemporary Word Dictionary” published by Jiyukokumin-sha[1, 2] as the textbase of contemporary words. This dictionary has many contemporary word entries and their descriptions. The notable features of the dictionary are as follows:

- The dictionary has about 9,000 contemporary word entries. Many newly-coined words are selected for inclusion every year.
- This dictionary is published every year and a large percentage of the previous year’s entries are replaced.
- The word entries have three upper level concepts. This means that these concepts and word entries construct a four-level thesaurus. The highest level concepts consist of 1) Economics and Management, 2) Information and Industry, 3) International Relations, 4) Politics, 5) Social Life, 6)

Health and Medicine, 7) Science and Technology, 8) Vogue and Fashion, 9) Sports and Hobbies, and 10) Culture and Art. These ten highest concepts are called “HighestConcepts” hereafter in this paper.

Outline of the Proposed System: Figure 1 shows an outline of the proposed system. The system has two major parts. One of them is a dictionary, $M_i(i = 1, 2 \dots n - 2, n - 1)$. In this paper, M_i is the same “Contemporary Word Dictionary” published by Jiyukokumin-sha. In Fig. 1, each $R_i(i = 1, 2, 3 \dots n - 1, n)$ is a candidate vogue word or concept for each year. R_{n-1} means words in vogue last year and R_{n-2} means words in vogue two years ago. R_n is words that will come into vogue in the future. Usually, R_n are created by the creator.

The aim of this research is to calculate the distance between a candidate vogue word and the social situation, i.e., the distance between R_i and M_{i-1} is calculated. Usually, R_i are newly-coined words and are not included in M_i .

In our experiment, $R_i(i = 1, 2, \dots n - 2, n - 1)$ are the “Most Popular Vogue Words,” selected by Jiyukokumin-sha every year. The “Most Popular Vogue Words” are selected every autumn. The number of selected words is about ten and these words are usually newly-coined words. Thus, the “Most Popular Vogue Words” in 1999 are not included in the Contemporary Word Dictionary 1999, because the 1999 dictionary is edited in the autumn of 1998. The words selected in 1999 should be predicted from the Contemporary Word Dictionary 1999.

3.2 CALCULATION OF DISTANCE

The proposed system has two major functions for predicting future vogue words or concepts. One is calculation of distance between a new candidate word and the Contemporary Word Dictionary. The other is trend analysis. The analysis shows that calculated distance is reliable in some HighestConcepts.

Distance Calculation for Candidate Word: The distance between the new word in R_i and the word entries in the dictionary M_{i-1} is calculated by using the “Vector Space Method,” which is well known in the natural language processing domain. To use the vector space method, each new word in R_i should have a description. The creator has to write this description. The system calculates the distances between the description of the new candidate word and descriptions for each word entry. A new candidate word having minimum distance is selected as the “best candidate.”

One problem of the vector space method is the sparseness of the vector space. The description of each word entry in the “Contemporary Word Dictionary” is very short, just several tens of words. On the other

hand, there are about 30,000 entries in the dictionary. The vector is thus very sparse, just 20 or 30 elements are non-zero in a vector whose size is about 30,000.

The calculated distances are strongly affected by the content of the description of the newly created word. This problem is experimentally and statistically analyzed in the following sections.

Trend Analysis: The other method of prediction is “trend analysis” using the Most Popular Vogue Words. The proposed system denotes the Most Popular Vogue Words as $R_i(i = 1, 2, 3 \dots n - 2, n - 1)$ ¹. All the Most Popular Vogue Words in $R_i(i = 1, 2, 3 \dots n - 2, n - 1)$ are different from each other. The system selects the best word for each R_i . We should be able to predict the future trend for vogue words from these calculation results.

The authors focused on word entries in the dictionary M_i that had a HighestConcept. The system selects minimum distance HighestConcept for each $R_i(i = 1, 2, 3, \dots n - 2, n - 1)$. In this case, each R_i has about ten words.

To solve the sparseness problem, we employed the 20 nearest neighbor approach. Twenty nearest word entries are selected for each newly created word². This means that the “kernel function” of the nearest neighbor method is of some size. In this case, the kernel function is “20 nearest neighbor.”

Usually, there are about ten Most Popular Vogue Words for each year. Thus, about $10 \times 20 = 200$ HighestConcepts are selected for one R_i . The number of HighestConcepts is 10. The percentage of each HighestConcept is calculated in this paper from about 200 selected nearest neighbors. For example, if the total number of selected HighestConcepts is equal to 200, and the number of “Politics” is 35, the calculated percentage is 0.175 (17.5%).

This percentage is calculated for all R_i , as shown in Fig. 2. The horizontal axis denotes year and the vertical denotes the calculated percentage, the value of which is normalized. The average percentage of the HighestConcept is equal to 1. The details of Fig. 2 are analyzed in the following section.

4. EXPERIMENTAL EVALUATION

4.1 EXPERIMENTAL RESULTS

The results shown in Figure 2 were calculated by using the 1999 CD-ROM of the Contemporary Word Dictionary. In this case, R_n is “Most Popular Vogue Words” in 1999. This R_n is assumed to be newly created candidate words. The R_n words are not included in the M_i dictionary. The purpose of the trend

¹Some words in $R_i(i = 1, 2, 3 \dots n - 2, n - 1)$ are included in the dictionary $M_i(i = 1, 2, 3 \dots n - 2, n - 1)$. These words are excluded from the analysis

²The “Most Popular Vogue Words” for $R_i(i = 1, 2, 3, \dots n - 2, n - 1)$.

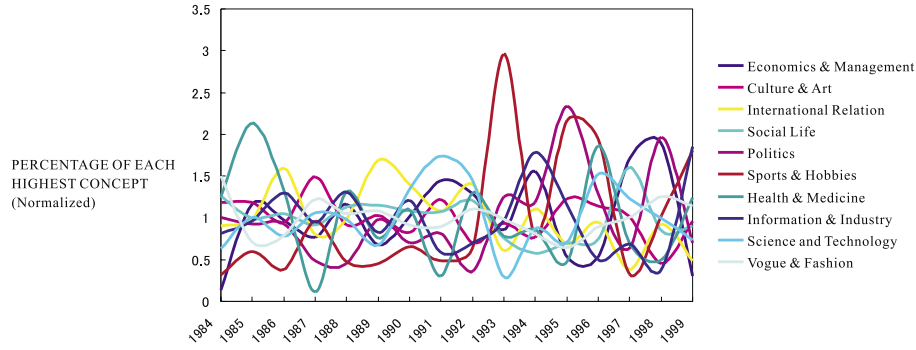


Figure 2: Trend Analysis by using 1999 Contemporary Word Dictionary

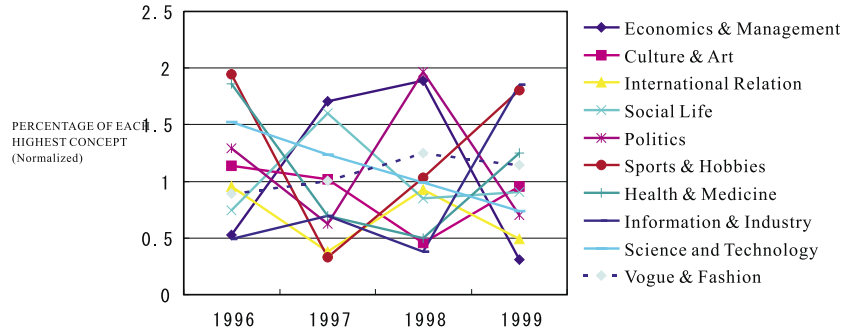


Figure 3: Detailed Trend of HighestConcept Percentage by using 1999 Contemporary Word Dictionary

analysis is to evaluate the newly created words or concepts, R_n , by using the “Most Popular Vogue Words” in $R_i (i = 1, 2, 3, \dots, n - 2, n - 1)$.

Figure 2 is very complex and it is difficult to read the vogue trend in the figure. However, the following analytical results can be derived.

- At the end of 1998, “Economics and Management” and “Social Life” have high percentages. But they are the peak points. This suggests that these two HighestConcepts will soon go down in rank.
- “International Relations” and “Sports and Hobbies” are increasing at the same time.
- “Culture and Art” and “Health and Medicine” are at a low level at the same time.

The results in Fig. 2 suggest that “International Relations,” “Sports and Hobbies,” “Culture and Art,” and “Health and Medicine” will have a powerful influence on 1999 vogue words and concepts.

Japanese TV stations broadcast two TV dramas in March, 2000. One was a love story about an “famous hairstylist” and a handicapped woman. This drama has two perspectives: “Culture and Art” and “Health and Medicine.” The other drama was also a love story.

In this drama, an international resistance fighter loves a woman. This beautiful woman is a typical “office lady”. In this case, there is only one perspective, “International Relations.”

The famous hairstylist drama got a very high program rating. The maximum value was 41.3%. On the other hand, the international drama had an average program rating of about 10%. This suggests that a combination of two HighestConcepts with increasing influence is a good tool to get a high program rating. This analytical result shows that it is possible for the proposed system to predict the future trend in vogue concepts.

4.2 EVALUATION OF VECTOR SPACE METHOD

One of the problems of the proposed vector space method is that the number of matched elements of this huge sized vector is very small, for instance, just three or four. Thus, the above result needs to be confirmed by other means.

The key means used to confirm the validity of the vector space result was human assessment of whether the matched elements (words) were reasonable. The data for this assessment were generated by performing the following steps.

Table 1: Best Five Matched Words for “Economics and Management” (Original words are in Japanese)

1990	Stock (9)	Year (6)	Bubble (5)	Economy (5)	Situation (4)
1991	Company (13)	Bond (12)	Stock (11)	Investor (10)	Price (9)
1992	Economy (15)	Bubble (12)	Collapse (11)	Year (7)	Depression (7)
1993	Price (4)	Bubble (3)	CEO (3)	Exercise (3)	Means (3)
1994	Price (13)	Commodities (9)	Falling (7)	Goods (5)	System (5)
1995	Year (3)				
1996					
1997	Bond (11)	Evolution (10)	Ban (9)	Big (9)	Trading (9)
1998	Banking-Organ (25)	Credit (23)	Bank (18)	Management (12)	Year (11)
1999					

Table 2: Best Five Matched Words for “International Relation” (Original words are in Japanese)

1990	President (10)	Iraq (9)	Year (8)	Forces (7)	World (7)
1991	Year (11)	Group (5)	Evolution (4)	Month (4)	Trouble (4)
1992	President (17)	Year (13)	Victory (9)	Election (9)	Nation (8)
1993	Japan (3)				
1994	Thought (5)	Society (4)	Policy (4)	Human (4)	Party (4)
1995	Party (6)	Election (5)	Candidate (3)		
1996	Prime Minister (6)	Month (4)	Politics (4)	Society (3)	English (2)
1997					
1998	Year (6)	Communist Party (5)	Prime Minister (5)	Election (4)	Committee (3)
1999	Reason (3)	Election (3)			

Table 3: RSI Values and the Number of Rising in the Next Year

RSI Values	Num. of Highest Concept	Number of Rising	Percentage (%)
0.1~0.19	1	1	100
0.2~0.29	2	2	100
0.3~0.39	16	10	62.5
0.4~0.49	30	13	43.3
0.5~0.59	35	10	26.6
0.6~0.69	14	4	28.6
0.7~0.79	2	0	0
0.8~0.89	0	0	0

- Count the appearance of each matched word for each year and HighestConcept.
- Eliminate the words whose count, n, is not more than two.
- Select the top five words for each HighestConcept and for each year.

Table 1 and Table 2 show the results for the HighestConcepts “Economics and Management” and “International Relations”, respectively. In these Tables, “(n)” shows the count. In Table 1, a popular word “Bubble” is observed for the years 1990, 1992, and 1993. This corresponds to the end of the “bubble economy” in Japan. “Big” and “Ban” in the year 1997 and “Banking-organ” and “Insolvency” in the year 1998

correspond to the point when Japanese banks were in severe economic crisis and the Japanese government changed banking policy. These results show that the matched words in the vector space method are reasonable based on current human knowledge of the social situation at that time. Thus, the calculated distance may be accurate and reliable for concept evaluation.

In Table 2, “Iraq” in the year 1990 and “(U.S) President” and “Election” in year the 1992 are reasonable. However, some words, for instance “Communist Party” in the year 1998 is not reasonable. This shows that calculated distance is less reliable for this HighestConcept. Generally speaking, it was clarified that the matched words were reasonable in the domains of politics and economics, but less reliable in those of “Vogue and Fashion,” “Sports and Hobbies,” and “Culture and Art.”

5. FINANCIAL ENGINEERING APPROACH

“Maximization of profit” and “diversification of risk” are major business aims and many theoretical approaches have been proposed in the financial domain. Those aims are also the focus of this paper. Thus, in this section, the RSI (Relative Strength Index) method for dealing in stocks and shares is applied to the trend analysis of the proposed system.

5.1 RSI: RELATIVE STRENGTH INDEX

RSI is a technical analysis method used for analyzing stock prices. The original definition is

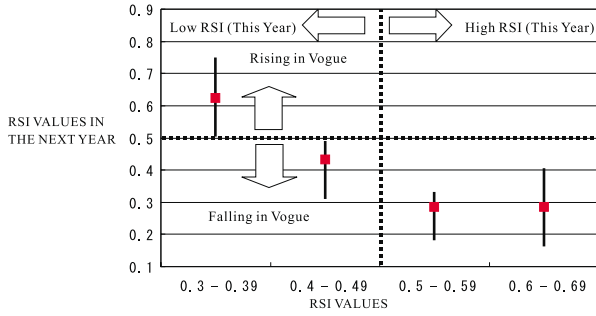


Figure 4: Bootstrap Result for Table 3

$$RSI = \frac{TPCforRising}{TPCforRising+TPCforFalling},$$

where,

TPCforRising = The total of price changes for the days whose price is rising,

TPCforFalling = The total of price changes for days whose price is falling.

In a stock market, the range of RSI calculation is two weeks (14 days). If the RSI value is less than 0.3, the stock sells at a low price. On the other hand, if the RSI value is more than 0.7, the stock sells at a high price. The RSI equation is very simple; the equation focuses in the curve or range of the estimated value. This approach is very similar to the proposed trend analysis method. In order to calculate the RSI value in the proposed system, the following equation is applied to the distance value shown in Fig. 2.

$$RSI = \frac{TNPCforRising}{TNPCforRising+TNPCforFalling},$$

where,

TNPCforRising = The total of normalized percentage changes for years whose percentage is rising,

TNPCforFalling = The total of percentage changes for years whose percentage is falling.

A range of six years is employed for the RSI calculation in the proposed system because otherwise the data are not sufficient. Table 3 shows the results. If the RSI value is small, the percentage of the HighestConcept rises in the following year. On the other hand, if the RSI value is high, the percentage falls in the following year. This result shows that RSI values provide good information to predict aconcept that will come into vogue in the future.

5.2 BOOTSTRAP TEST FOR RSI VALUES

The RSI values in Table 3 also need to be tested statistically. Traditional statistical test methods cannot be applied because RSI is a non-linear function. Thus, the “Bootstrap method” proposed by Efron and Tibshirani[3] was applied. A notable feature of the

Bootstrap method is “repeated re-sampling from the identical examples.” The statistical RSI distribution calculated from the re-sampled examples are assumed to be equal to that of real examples in the Bootstrap approach.

Generally speaking, the following conditions should be met when using the Bootstrap method.

- The number of samples should be more than nine. If the number is less than ten, the calculation result is not reliable[4].
- The number of repetitions of re-sampling from the identical examples is 10,000 for the standard deviation calculation.

Figure 4 shows the Bootstrap results. The black bar shows the one-sigma intervals calculated using the Bootstrap method. The black bar does not include the value zero. This means that RSI values are expressive at 95% probability. HighestConcepts having very low RSI values will become vogue concepts in the near future.

6. CONCLUSION

This paper has demonstrated a new system using natural language processing techniques for the evaluation of a word or concept in vogue. The distance between a newly created word and the word entries in the Contemporary Word Dictionary is calculated by using the well-known Vector Space Method. Experimental results show that the future trend can be predicted with a certain measure of success, especially in Economics and Management and “International Relations”, which are two of the highest concepts of that dictionary. The validity of the system’s performance was confirmed statistically. However, the distance calculation is not expressive in the other highest concepts of the dictionary. The major reason for this is the sparseness of the vectors in the Vector Space Method. The combination of the Contemporary Word Dictionary and a general Japanese language dictionary should be investigated in further research.

REFERENCES

- [1] “Contemporary Word Dictionary 1999 (In Japanese)”, Jiyukokuminsha, Tokyo, 1998.
- [2] “Contemporary Word Dictionary 2000 (In Japanese)”, Jiyukokuminsha, Tokyo, 1999.
- [3] B. Efron and R. J. Tibshirani, “An Introduction to the bootstrap,” Chapman & Hall, 1993.
- [4] M. R. Chernick, “Bootstrap Methods – A Practical Guide,” Chap. 9. in “Too Small Sample Size,” John Wiley & Sons, Inc. 1999.