An Analysis of Users in a Q&A Site Submitted Many Answers Where First Polar Words are Negative Words

Masashi Minamiguchi, Kenji Umemoto, Yasuhiko Watanabe, Ryo Nishimura, Yoshihiro Okada
Department of Media Informatics, Ryukoku University
Seta, Otsu, Shiga, Japan
Email: t12m107@mail.ryukoku.ac.jp, t11m074@mail.ryukoku.ac.jp,
watanabe@rins.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp, okada@rins.ryukoku.ac.jp

Abstract—In this study, we investigate that the evaluations of answers in a Q&A site are affected by whether the first polar words are negative words. We first investigate answers where the first polar words were negative words. The result shows that the evaluations of answers in a Q&A site were less affected by whether the first polar words were negative words. Furthermore, we investigate users in a Q&A site who submitted many answers where the first polar words were negative words. The result show that answers submitted to a Q&A site were evaluated based on whether their explanations were detailed and informative, rather than whether the first polar words were negative words. In this study, we use the data of Yahoo! chiebukuro, a widely-used Japanese Q&A site, for observation and examination. We also use the opinion extraction tool and model data, which were developed by Knowledge Clustered Group in National Institute of Information and Communications Technology (NICT).

Keywords—negative word; opinion polarity; opinion expression; Q&A site; Yahoo! chiebukuro.

I. INTRODUCTION

Some words have the polarity. These words are called *polar words* and classified into positive words and negative words [1]. In face-to-face communication, some people tend to speak stories where the first polar words are negative words, even if their stories include not only negative words but also positive words. This way of speaking stories may affect the evaluations of their stories. Furthermore, negative non-verbal signals, such as folded arms, frowning, distance increase, and looking away, often make their stories more negative. On the other hand, in question and answer (Q&A) sites, such as Yahoo! answers [2], Yahoo! chiebukuro [3], Oshiete! goo [4], Hatena [5], and OKWave [6], negative non-verbal signals rarely make their stories more negative. It is because it is difficult to send negative non-verbal signals via Q&A sites. However, we do not know whether the way of writing stories where the first polar words are negative words affects the evaluations of their stories, especially, answers submitted to Q&A sites. To solve this problem, we investigate answers where the first polar words are negative words. Furthermore, we investigate users who submitted many answers where the first polar words were negative words. In this study, we used the data of Yahoo! chiebukuro, a widely-used Japanese Q&A site, for observation and examination.

The rest of this paper is organized as follows: In Section II, we surveys the related works. In Section III, we describes the data of Yahoo! chiebukuro, which we used for observation and examination. In Section IV, we describes a method of detecting users submitted many answers where the first polar words were negative words. In Section V, we show the experimental results and discussions. Finally, in Section VI, we present our conclusions.

II. RELATED WORKS

In these years, a large number of studies have been made on sentiment analysis based on the polarities of words. Sharifi and Cohen proposed a method of using conditional random fields for extracting polar words and determining the overall sentiment of text [1]. Takamura et al. developed a lexical network out of glosses in a dictionary, a thesaurus and a corpus, and extracted the semantic polarities of words by regarding semantic polarities of words on the network as spins of electrons [7]. Goto et al. proposed a method for improving the performance of the polarity lexicon extraction based on Takamura et al.'s spin model [8]. Ikeda et al. propose a machine learning based method of sentiment classification of sentences by using the polarities of words [9]. Also, in these years, a large number of studies have been made on how to evaluate answers submitted into Q&A sites. Kuriyama et al. proposed a method of evaluating answers by using answerers' records of postings to a Q&A site [10] [11]. Ishikawa et al. proposed a method of evaluating answers by using machine learning techniques [12]. However, there are few studies whether the evaluation of answers are affected by the order of polar words. On the other hand, Kido reported that rhetorical structure is affected by the order of non-fact sentence (e.g., comments and opinion sentences) while rhetorical structure is less affected by the order of fact sentence (e.g., sentences that report actual survey figures) [13]. However, Kido did not consider the order of polar words in reports.

III. YAHOO! CHIEBUKURO

In this study, we used the data of Yahoo! chiebukuro for observation and examination. In Japanese, chiebukuro means "bag of wisdom". The data of Yahoo! chiebukuro was published by Yahoo! JAPAN via National Institute of Informatics in 2007 (http://research.nii.ac.jp/tdc/chiebukuro.html). This data consists of about 3.11 million questions and 13.47 million answers, which were posted on Yahoo! chiebukuro

TABLE I.   THE NUMBERS OF USERS AND THEIR SUBMISSIONS TO PC CATEGORY, SOCIAL ISSUES CATEGORY, AND ALL 286 CATEGORIES IN YAHOO! CHIEBUKURO (FROM APRIL/2004 TO OCTOBER/2005).

category	number of questions	number of questioners	number of answers	number of answerers
PC	171848	43493	474687	27420
social issues	78777	13259	403306	25766
all 286 categories	3116009	165064	13477785	183242

TABLE II.   THE EXTRACTION RESULT OF ANSWERS WHERE THE FIRST POLAR WORDS WERE NEGATIVE WORDS (SOCIAL ISSUE CATEGORY IN THE DATA OF YAHOO! CHIEBUKURO).

answers	number of answers	best answer ratio
all the answers in social issue category	403306	19.5
answers where the first polar words were negative words	162878	19.5

from April/2004 to October/2005. In the data, each question has at least one answer because questions with no answers were removed. Each user can submit his/her answer only one time to one question. Each questioner is requested to determine which answer to his/her question is best. The selected answer is called the *best answer*. In order to avoid identifying individuals, user accounts were replaced with unique ID numbers. By using these ID numbers, we can trace any user's questions and answers in the data. Table I shows the numbers of users and their submitted messages (questions and answers) to PC category, social issues category, and all 286 categories in the data.

In order to detect answers which include negative words, we use an opinion extraction tool [14]. This tool was developed by Knowledge Clustered Group in National Institute of Information and Communications Technology (NICT) and released in September/2011. This tool detects opinion expressions in given sentences and outputs the polarities of them (positive/negative polarity) when they have the polarities. Knowledge Clustered Group in National Institute of Information and Communications Technology (NICT) also developed and released model data for the opinion extraction tool [15]. This model data consists of about 35000 words (10000 positive words and 25000 negative words). These words were extracted from 20000 sentences in Web document corpus. This model data is useful to detect negative words in answers precisely. For example, when we apply this tool to (A1), the opinion extraction tool detect negative words in second sentence "*jimin tou nado ga matomo na kyougi, touron mo naku kyoukou ni kokki kokka wo kimeta.* (The Liberal Democratic Party and other political parties set Kimigayo as Japan's national anthem, without sufficient discussions in the Diet.)" and third sentence "*konna daiji na koto wo jibun tachi no omou toori ni kyoukou shita noda.* (It was so serious, however, they got their way.)".

(Q1)   *Hinomaru kimigayo wo kyohi suru hito ni shit-sumon desu. puro yakyu nado no kansen no toki mo yahari kyohi desu ka? K-1 nado ha takoku no kokka ni tsuite mo kiritsu wo motome rare masu ga donoyouna kanngae de dou koudou suru no desu ka?* (I have a question to persons who deny Hinomaru and Kimigayo. Do you deny them even when you watch professional baseball games? In case of sport games like K-1, we are asked to

stand up during singing of the national anthem, not only ours but other's national song. Tell me what you think and how you act.)

(A1)   *touzen kyohi desu. jimin tou nado ga matomo na kyougi, touron mo naku kyoukou ni kokki kokka wo kimeta. konna daiji na koto wo jibun tachi no omou toori ni kyoukou shita noda. motto shintyo ni kimeru hitsuyou ga atta.* (Of course, I refuse. The Liberal Democratic Party and other political parties set Kimigayo as Japan's national anthem, without sufficient discussions in the Diet. It was so serious, however, they got their way. We should discuss this issue more careful.)

Both (Q1) and (A1) were submitted to social issues category in Yahoo! chiebukuro. (A1) was an answer to (Q1). In case of (A1), the opinion extraction tool determines the second sentence is the first sentence which include an opinion expression, and the polarity is negative. As a result, (A1) is determined to be an answer where the first polar word was a negative word. By using this opinion extraction tool, we extract answers where the first polar words were negative words from the data of Yahoo! chiebukuro.

We applied this tool to 403306 answers submitted to social issues category in Yahoo! chiebukuro, and obtained 162878 answers where the first polar words were negative words. Table II shows the result of this extraction. As shown in Table II, in social issue category, the best answer ratio of answers where the first polar words were negative words is similar to that of all the answers. As a result, it may be said that the evaluations of answers in Yahoo! chiebukuro were less affected by whether the first polar words are negative words.

IV.   USERS SUBMITTED MANY ANSWERS WHERE THE FIRST POLAR WORDS WERE NEGATIVE WORDS

In this study, we detect users who submitted many answers where the first polar words were negative words to Yahoo! chiebukuro and investigate whether they got good evaluations of their answers. In order to detect and discuss users who submitted many answers where the first polar words were negative words, we test one hypothesis, Hypothesis NWFA:

TABLE III.　THE RELATION BETWEEN THE BEST ANSWER RATIO AND THE AVERAGE NUMBER OF SENTENCES

	the average number of sentences		
	1.0 – 3.0	3.0 – 6.0	6.0 –
number of users	8	13	12
best answer ratio of answers where the first polar words were negative words	21.7	25.3	33.3

Hypothesis NWFA　　If user i submitted not many answers where the first polar words were negative words, we would expect that user i submitted at most $N_{NWFA}(i)$ answers where the first polar words were negative words.

$$N_{NWFA}(i) = P_{NWFA} \times ans(i) \qquad (1)$$

where $ans(i)$ is the number of user i's answers and P_{NWFA} is the probability that one randomly selected answer is an answer where the first polar word is a negative word. As a result, P_{NWFA} is

$$P_{NWFA} = \frac{N_{NWFA}}{N_{ans}} \qquad (2)$$

where N_{ans} is the number of all the answers and N_{NWFA} is the number of answers where the first polar words are negative words. As shown in Table II, N_{ans} and N_{NWFA} of social issue category in Yahoo! chiebukuro are 162878 and 403306, respectively. As a result, P_{NWFA} of social issue category is 0.404.

When this hypothesis is rejected by a one-sided binomial test, we determine that user i submitted many answers where the first polar words are negative words.

V.　EXPERIMENTAL RESULTS AND DISCUSSIONS

We applied the detection method based on Hypothesis NWFA to 25766 users who submitted one or more answers to social issue category in Yahoo! chiebukuro. In this study, the significant level of Hypothesis NWFA was set to 0.000000005. It was extremely low because we intended to detect users who submitted extremely many answers where the first polar words were negative words. Our method detected 33 users and the best answer ratio of their answers is 25.4%. It is higher than the best answer ratio of all the answers submitted to social issue category in Yahoo! chiebukuro (19.5%). In face-to-face communication, persons who tend to speak stories where the first polar words are negative words often get poor evaluations of their stories. On the other hand, in Q&A sites, answerers who tend to write answers where the first polar words are negative words often get good evaluations of their answers. As a result, it may be said that answers submitted to Q&A sites were evaluated based on whether explanations were detailed and informative, rather than whether the first polar words were negative words.

Next, we discuss the relation between the best answer ratio and the average number of sentences in these 33 users' answers. We classified the detected 33 users into three groups depending on the average number of sentences in their answers indicated below:

group A	less than 3.0
group B	not less than 3.0 and less than 6.0
group C	not less than 6.0

Table III shows the number of users in each group and the best answer ratio of their answers where the first polar words were negative words. As shown in Table III, the best answer ratio is often high when the average number of sentences in answers is large. It is because, we think, explanations consisted of many sentences is often more detailed and informative than those consisted of few sentences. As a result, it also may be said that answers submitted to Q&A sites were evaluated based on whether explanations were detailed and informative, rather than whether the first polar words were negative words. For example, user 273731 submitted 203 answers where the first polar words were negative words to social issue category. Because the average number of sentences in his/her answers was 6.7, user 273731 was classified into group C. The explanation of user 273731's answers were generally detailed and informative. Possibly because of it, the best answer ratio of user 273731's answers is 45.3 %. It is higher than the average of the best answer ratio of answers submitted into social issue category (19.5%). On the other hand, user 169784 submitted 131 answers where the first polar words were negative words to social issue category. Because the average number of sentences in his/her answers was 2.8, user 169784 was classified into group A. The explanation of user 169784's answers were generally short and not informative. Possibly because of it, the best answer ratio of user 169784's answers is 15.6 %. It is lower than the average of the best answer ration of answers submitted into social issue category (19.5%).

VI.　CONCLUSION AND FUTURE WORK

In this study, we investigated answers where the first polar words were negative words and found that the evaluations of answers in Yahoo! chiebukuro were less affected by whether the first polar words were negative words. Furthermore, we investigated users who submitted many answers where the first polar words were negative words to Yahoo! chiebukuro, and found that these users often get good evaluations of their answers in Yahoo! chiebukuro. We think that answers submitted to Q&A sites were evaluated based on whether explanations were detailed and informative, rather than whether the first polar words were negative words.

In the future, we intend to investigate whether the evaluations of answers in Yahoo! chiebukuro are affected by whether the first polar words are positive words. Furthermore, we want to investigate various kinds of online documents, for example, messages in blog comments, web-based bulletin boards, and micro blogs.

R EFERENCES

[1] M. Sharifi and W. Cohen, "Finding domain specific polar words for sentiment classification," in *the LTI Student Research Symposium*, 2008. [Online]. Available: http://www.cs.cmu.edu/~mehrbod/polarity_08.pdf [retrieved: May, 2013]

[2] *Yahoo! Answers*, Yahoo!, 2005. [Online]. Available: http://answers.yahoo.com/ [retrieved: May, 2013]

[3] *Yahoo! chiebukuro*, Yahoo! JAPAN, 2004. [Online]. Available: http://chiebukuro.yahoo.co.jp/ [retrieved: May, 2013]

[4] *Oshiete! goo*, NTT Resonant Incorporated, 2000. [Online]. Available: http://oshiete.goo.ne.jp/ [retrieved: May, 2013]

[5] *Hatena*, Hatena Co., Ltd., 2005. [Online]. Available: http://q.hatena.ne.jp/ [retrieved: May, 2013]

[6] *OKWave*, OKWave, 2000. [Online]. Available: http://okwave.jp/ [retrieved: May, 2013]

[7] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientations using spin model," in *Transactions of Information Processing Society of Japan*, vol. 47, no. 2, Feb. 2006, pp. 627–637.

[8] T. Goto, Y. Kabashima, and H. Takamura, "Extracting semantic orientations using lexical networks: Performance improvement from the viewpoint of statistical mechanics," in *Technical Report of The Institute of Electronics, Information and Communication Engineers (IEICE) on Information-Based Induction Sciences and Machine Learning (IBISML)*, vol. 110, no. 265, Oct. 2010, pp. 19–25.

[9] D. Ikeda, H. Takamura, and M. Okumura, "Learning to shift the polarity of words for sentiment classification," in *Transactions of the Japanese Society for Artificial Intelligence*, vol. 25, no. 1, Jan. 2010, pp. 50–57.

[10] K. Kuriyama and N. Kando, "Analysis of questions and answers in q&a site (2) - based on document structures and attributes -," in *Technical Report of Information Processing Society of Japan (IPSJ)*, vol. 2009-FI-96, no. 3, Nov. 2009, pp. 1–8.

[11] K. Kuriyama and N. Kando, "Analysis of questions and answers in q&a site (3) - predicting best-answers based on users' histories -," in *Technical Report of Information Processing Society of Japan (IPSJ)*, vol. 2010-FI-97, no. 7, Jan. 2010, pp. 1–8.

[12] D. Ishikawa, T. Sakai, Y. Seki, K. Kuriyama, and N. Kando, "Automatic prediction of high-quality answers in community qa," in *Joho Chishiki Gakkaishi (Japan Society of Information and Knowledge)*, vol. 21, no. 3, Sep. 2011, pp. 362–382.

[13] M. Kido, "The influence of sentence ordering and sentence function on rhetorical structure : the ordering of fact and non-fact sentence," in *University of Tshukuba International Student Center Journal of Japanese language teaching*, vol. 12, Feb. 1997, pp. 1–10.

[14] *Opinion extraction tool (version 1.2)*, Knowledge Clustered Group of National Institute of Information and Communications Technology (NICT), 2012. [Online]. Available: http://alaginrc.nict.go.jp/opinion/ [retrieved: May, 2013]

[15] *C-3 Model for opinion extraction tool (version 1.2)*, Knowledge Clustered Group of National Institute of Information and Communications Technology (NICT), 2012. [Online]. Available: http://alaginrc.nict.go.jp/opinion/ [retrieved: May, 2013]