International Journal of Engineering and Technology Innovation, vol. 6, no. 2, 2016, pp. 165 - 172

Statistical Analysis of Automatic Seed Word Acquisition to Improve Harmful Expression Extraction in Cyberbullying Detection

Suzuha Hatakeyama[1,*], Fumito Masui[1], Michal Ptaszynski[1], Kazuhide Yamamoto[2]

[1] Department of Computer Science, Kitami Institute of Technology, Kitami, Japan.

[2] Department of Electrical Electronics, and Information Engineering, Nagaoka University of Technology, Nagaoka, Japan.

Received 15 January 2016; received in revised form 12 March 2016; accepted 25 March 2016

Abstract

We study the social problem of cyberbullying, defined as a new form of bullying that takes place in the Internet space. This paper proposes a method for automatic acquisition of seed words to improve performance of the original method for the cyberbullying detection by Nitta et al. [1]. We conduct an experiment exactly in the same settings to find out that the method based on a Web mining technique, lost over 30% points of its performance since being proposed in 2013. Thus, we hypothesize on the reasons for the decrease in the performance and propose a number of improvements, from which we experimentally choose the best one. Furthermore, we collect several seed word sets using different approaches, evaluate and their precision. We found out that the influential factor in extraction of harmful expressions is not the number of seed words, but the way the seed words were collected and filtered.

Keywords: cyberbullying, information extraction, text mining, seed word, SO-PMI-IR

1. Introduction

The method by Nitta et al. [1] uses a Web mining technique (with Yahoo! API) to extract co-occurrences of input sentences with cyberbullying entries and calculates relevance values using a modified SO-PMI-IR formula [4]. We repeated their experiment from 2013, with exactly the same settings, to find out how the performance of the Web-based method changed during the several years since being proposed in 2013 and compare it to the results reported by Nitta et al. As the evaluation metrics, we used area under the curve (AUC) on the graph showing precision and recall. The result of the comparison is represented in Fig. 1.

Nitta et al. [1] report a high performance (91%) when it comes to precision for low recall window. Moreover, originally the performance of their method retained high performance (around 70%) even close to 50% of recall, after which it decreased overall performance due to drop in precision for higher thresholds. However, when we repeated their experiment in January 2015, the overall results greatly dropped. Precision did not exceed 60% for the whole threshold span.

In the following sections, we firstly discuss the potential reasons for the drop in performance (Section 2). Next we explain the original method (Section 3) and then propose a method for automatic acquisition of seed words to improve the original method (Section 4). Next, we conduct a series of experiments with automatically obtained different seed word sets, experimentally choose the one with the best performance (Section 5) and discuss the results (Section 6). Finally we conclude the paper (Section 7).

* Corresponding author. E-mail address: hatakeyama@ialab.cs.kitami-it.ac.jp
Tel.: +81-0157-26-9349

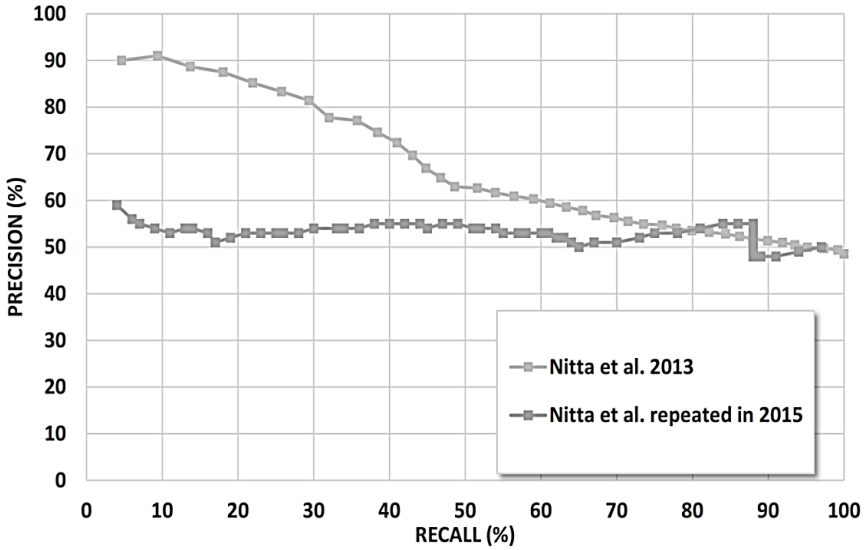

Fig. 1 Performance comparison for method by Nitta et al. (2013) reported originally in 2013 and repeated in 2015

2. Discussion on Reasons for Performance Decrease

In this section we discuss possible reasons for the decrease of the performance in the method proposed by Nitta et al. [1].

After thorough analysis of the experiment data, we noticed that most of the information extracted in 2013 was not available in 2015. This could be due to the following reasons. Firstly, fluctuation in page rankings could push the information to the bottom which makes it not extractable anymore by Nitta et al.'s method. Secondly, frequent deletion requests of harmful contents by Internet Patrol members could make their efforts pay off.

However, the most probable is the third cause, which is the recent tightening of usage policies by most Web service providers, such as Google[*], Twitter[†] and Yahoo!, which are used by Nitta et al. This includes recently introduced DMARC[‡] policies related to e-mail spoofing and general improvements in policies aimed at decreasing Internet harassment.

The fact that the performance of Nitta et al.'s method decreased from over 90% to less than 60% during less than 3 years is an important warning also for other research based on Web mining. Probabilities of such problems have been indicated some time ago [2], and could become a major problem in the future. This also advocates more focus on either corpus-based methods or methods providing constant update of applicable seed words (features), such as the one proposed in this paper.

*kiero (get lost), shineyo (just die!), munou (incompetent), busaiku (ugly), busu (ugly hag), doutei (virgin), chon (f*ckin' Chinese), BUSAIKU (ugly), shine (die), uzai (annoying), UZAi (ANNOY- ing), KIMOi (UGLy), kichigai (fucking crazy), kimoi (ugly), kasu (scum), korosu (kill), kuzu (trash)*

Fig. 2 Examples of extracted harmful statements. Japanese Romanized (J), English translation (E), adjusted to express the original meaning as closely as possible

3. Description of Original Method

The method originally proposed by Nitta et al. [1] uses a list of seed words to calculate semantic orientation score SO-PMI-IR and then maximize the relevance of categories with input sentence. There are three steps in the classiftcation of the harmfulness of input:

(1) Phrase extraction,

(2) Categorization and harmful word detection together with harmfulness polarity de- termination,

(3) Relevance maximization.

This method is an extension of the method proposed by Turney [4] to calculate the relevance of words with specified categories according to the equation 1, where p_i is a phrase extracted from the input, w_j are three words that are registered in one category of harmfulness polarity words, hits(p_i) and hits(w_j) are Web search hits for each category for p_i and w_j respectively, hits(p_i&w_j) is a number of hits when p_i and w_j appear on the same Web page. Finally, $PMI - IR(p_i, w_j)$ is the relevance of p_i and w_j.

$$\text{SO-PMI-IR}(p_i,w_j) = \log_2\left\{\frac{hits(pi\&wj)}{hits(pi)hits(wj)}\right\} \quad (1)$$

Turney's method was extended by Nitta et al. to work not only on words, but on phrases. The phrases are automatically extracted from input sentences using depen- dency relations. Next, for all of the phrases the relevance is calculated with seed words from multiple categories of harmful words. The degree of association for each category, SO-PMI-IR score is maximized so that the maximum value achieved within all mea- sured categories is considered as the harmfulness *score* representative for the input sentence, and is calculated according to the equation 2.

$$\text{score} = \max(\max(\text{SO-PMI-IR}(p_i,w_j))) \quad (2)$$

4. Existing Mechanism

In this section we propose our reimplementation and performance improvement of the method for cyberbullying detection. This method exists for automatic acquisition and update of seed words used in Nitta et al.'s method.

4.1. Automatic Acquisition of Harmful Words

To improve the performance of Nitta et al.'s method we apply a method proposed by Ishizaka et al. [3].

The method proposed by Ishizaka et al. extends the initial value based on n-gram statistics and uses it to automatically acquire harmful expressions from a corpus (for example the Web). It calculates a SO-PMI-IR value (Semantic Orientation mining method based on Point Mutual Information with the statistical data collected by Information Retrieval [4]; here we abbreviate this to "SO value".) indicating the degree of harmfulness of a word in a document. If SO value for a word is greater than or equal zero, the word is considered harmful. The method also allows controlling the extraction precision by changing the threshold window of n-gram co-occurrence probability.

Originally, the method was successfully applied to harmful word acquisition from a large Japanese electronic bulletin board system (BBS), called "2-channel." Some of the examples of harmful statements extracted with this method are represented in Fig. 2. All example sentences contain harmful expressions.

The method of Ishizaka et al. allows modification and updating of the initial value in Nitta et al.'s method, which consists of a set of harmful seed words used in the IR procedure (Web mining). Such words could be automatically acquired and updated regularly, which would greatly decrease the cost of collecting the seed word lists when the Web mining conditions change.

4.2. Double Filtering of Harmful Seed Words

We conducted an experiment using Ishizaka et al.'s method to verify its potential in automatic selection of seed word candidates for Nitta et al.'s cyberbullying detection method.

Ishizaka et al. at first conducted a questionnaire in which they asked three annotators to judge whether a word is harmful or not. In the questionnaire, they used a set of 2,735 words automatically extracted from harmful BBS entries. From the initial set, they retained only those words for which all three annotators agreed about. There were 76 of such words (we refer to them as harmful word candidates). These words were used as a base for automatic selection. Next, Ishizaka et al. automatically selected a set of 14 most frequent harmful and 17 non-harmful basic words appearing in the original entries by using n-gram statistics (with n=7). They used those basic words to filter-out he most harmful out of the initial 76 harmful word candidates selected by human annotators. In our approach we added to the Ishizaka et al.'s 14 harmful basic words additional 9 harmful basic words used by Nitta et al. With overall 23 harmful and 17 non-harmful basic words we conducted first filtering of the basic words.

We calculated SO values for each of the 76 harmful word candidates with every other basic word, and calculated the sum of SO values for each word candidate. From the words which obtained SO value higher than 0 we selected those words which also appeared as basic words. The basic word set used in filtering is represented in Table 1. The basic words that were filtered out are represented in Fig. 3. The first filtering provided a set of basic harmful words which were both harmful from the point of view of human annotators and were in the strongest relevance to harmful BBS entries.

Table 1 Harmful and non-harmful basic words used in the experiment, with their English translations (meaning as close to original as possible; capital letters indicate different Japanese characters used to write a word)

	Harmful basic word candidates	Non-harmful basic word candidates
Ishizaka [3]	*kiero* (get lost), *ujimushi* (worm), *kasu* (scum), *shineyo* (just die!), *doutei* (virgin), *UZAi* (ANNOYing), *KIMOi* (UGLy), *kichigai* (fucking crazy), *chon* (f*ckin' Chinese), *kuzu* (trash), *munou* (impotent), *BUSAIKU* (ugly), *busu* (ugly hag), *kirai* (hate)	*hikikae* (exchange), *kaiage* (purchase), *shiborikomi* (narrowing), *kawaii* (cute), *furikae* (transfer), *koujun* (descending), *suteki* (nice), *ikemen* (handsome), *akai* (red), *subarashii* (wonderful), *utsukushii* (beautiful), *tsukue* (desk), *suppai* (sour), *chuurippu* (tulip), *taiyou* (sun), *natsu* (summer), *shikakui* (square)
Nitta [1]	*sekkusu* (sex), *yariman* (slut), *fera* (bl*wjob), *shine* (die), *korosu* (kill), *naguru* (beat the crap out of somebody), *uzai* (annoying), *kimoi* (gross), *busaiku* (ugly)	

1. J : *Busaiku dashi uta wa heta dashi dansu mo heta.*
   E : She's ugaly, sings badly, and cannot even dance.
2. J : *Omake ni DQN darake.*
   E : Apart form that all moron.
3. J : *Kumamon ni kuwasero, baka bono!*
   E : Feed this to Kumamon [TV character], you ass*ole!
4. J : *Ramen mania tte shinsoko kimoi*
   E : That Ramen maniac is like completely gross
5. J : *Munou dakara darou.*
   E : That's 'cause his an impotent.

Fig. 3 Harmful basic words (SO value ≥ 0)

In the second filtering we used as basic words: the 17 harmful basic words obtained in the first filtering and the same 17 non-harmful basic words (later called "seed-17"), and calculated the SO value again, this time with a dataset containing 76 harmful word candidates and 17 non-harmful basic words (93 words in total). We added the non-harmful words to verify whether the harmful basic words cause a bias in scoring. If so, the non-harmful words would also achieve a high SO value.

For a comparison, we also used only Nitta's original 9 harmful seed words instead of the filtered out 17 basic words (later called "seed-9"), and similarly calculated the SO value for the whole dataset (93 words). Having all SO values, we calculated arithmetic mean SO μ and standard deviation σ and automatically set a threshold $\alpha 1$ as $\mu \pm \sigma$.

Finally, we extracted candidates for harmful words *w* and non-harmful word candidates *h* as $h \leq \alpha 1 \leq w$. A graph visualizing this operation is represented in Fig. 4. All words extracted this way for sets seed-17 and seed-9 are represented in Table 2. Worth mentioning is the fact that although one would expect the basic words to come on top, they would be expected to obtain higher SO value due to perfect relevance score with themselves, resulting in the candidate words extracted according to this procedure did not contain even one word from the basic harmful word set. This indicates that the double filtering optimized the search and can be considered a viable procedure.

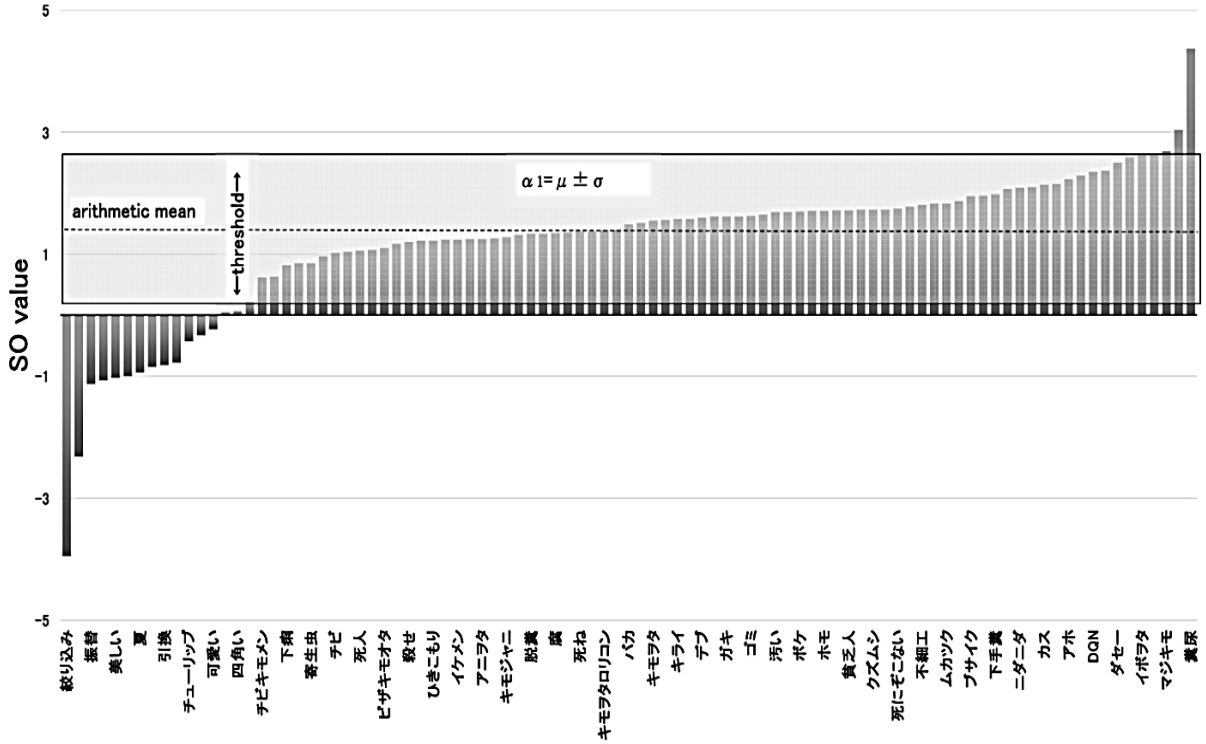

Fig. 4 Graph visualizing the procedure for extraction of word candidates for set seed17

Table 2 Harmful and non-harmful seed words obtained according to the double filtering procedure (seed-17) and for original words used by (Nitta et al. 2013) (seed-9)

	Harmful seed words	Non-harmful seed words
Seed-17	*dasee* (hopeless), *bakasayo* (stupid communist), *fun-nyou* (dung), *majikimo* (seriously gross), *ibo-ota* (warty nerd), *kuzumasugomi* (shitty mass-mudia), *goki-ota* (cockroachy nerd)	*shiborikomi* (narrowing), *kaiage* (purchase), *furikae* (transfer), *koujun* (descending), *natsu* (summer), *suteki* (nice), *utsukushii* (beautiful), *kawaii* (cute), *hikikae* (exchange), *akai* (red), *taiyou* (sun), *chuurippu* (tulip), *tsukue* (desk)
Seed-9	*bicchi* (bitch), *ibo-ota* (warty nerd), *mekuso* (eye gum), *dappun* (feces), *kusomushi* (dung beetle), *bakasayo* (stupid communist), *dasee* (hopeless), *fun-nyo* (dung), *goki-ota* (cockroachy nerd), *maji-kimo* (seriously gross), *kuzumasugomi* (shitty mass-mudia), *gumin* (ignorant)	*shiborikomi* (narrowing), *chuurippu* (tulip), *kaiage* (pur-chase), *suteki* (nice), *taiyou* (sun), *tsukue* (desk), *natsu* (summer), *akai* (red), *shikakui*(square), *koujun* (descending), *kawaii* (cute), *hikikae* (exchange), *furikae* (transfer), *kiseichuu* (parasite), *utsukushii* (beautiful)

5. Experiment and Results

After obtaining seed word sets described in section 3, we conducted an experiment to verify whether and how much does the change in the selection of seed words affect Nitta's method. Moreover, we calculated the statistical significance of the results by using the McNemar Test [5].

The experiment was designed as follows. Using filtered seed words we prepared six different harmful word candidate sets (later referred to as "cases"), applied them as seed words in Nitta's method and verified its performance. We show 6 cases of seed word combinations in Table 3.

Table 3 Description of six seed word sets (case 1– case 6) applied in Nitta's method in the experiment

case 1	7 words from seed17 (Table2)
case 2	12 words from seed9 (Table2)
case 3	16 words (7 words from seed-17 + Nitta's original 9 words)
case 4	21 words (12 words from seed-9 + Nitta's original 9 words)
case 5	5 words originally used by Ishizaka et al. [3]
case 6	9 words originally used by Nitta et al. [1]

Case 1 contained 7 double-filtered harmful words (all harmful words under seed-17 in Table 2). Case 2 contained 12 harmful words obtained by single-filtering out the words with the highest harmful relevance using only Nitta et al.'s original seed word set (all words under seed-9 in Table 2). Case 3 contained Nitta's original 9 seed words and 7 double-filtered harmful word candidates (overall 16 words). Case 4 contained Nitta's original 9 seed words and 12 single-filtered candidates. Case 5 contained 5 words originally used by Ishizaka in their article [3]. Finally, case 6 contained only Nitta's original 9 seed words without any modification.

In the evaluation, we looked at several metrics to choose the best seed word set. We calculated precision (P), recall (R), and balanced F-score (F) for each case. Then, we first (1) checked which case obtained the highest F-score for the longest threshold span. Next, (2) we looked at which case obtained the highest break-even point (BEP) on a plot with P and R. Thirdly, (3) we looked at which of the cases obtained the highest precision in general. Finally, (4) as the last mean of evaluation we used the same metrics as Nitta et al. [1], namely, area under the curve (AUC) on a plot with P and R. We also asked ourselves a question (5) whether it was always better to simply add more words, or whether the automatic filtering of harmful word candidates improved the method and to what extent. Below we address these five questions one by one.

(1) The highest F-score in general was obtained by case 2 (68%). Also case 5 and case 6 obtained similarly high F-score once, however, for most of the threshold, F for case 2 was always the highest, while for case 5 and case 6 mostly the lowest.

(2) The highest BEP was achieved in order by case 2 (0.59), case 1 (0.56), case 3 (0.55), case 6 (0.54), case 5 (0.53), and case 4 (0.52). This again suggests cases using seed word filtering as better then the others with case 2 achieving the highest score. The values of BEP for precision and recall for cases 1 to 6 are represented in Table 4.

(3) The highest P was obtained by case 4 (80%). However, its precision, although obtaining the highest score once, quickly decreases, ending with one of the worst results in general. The second best P was obtained by case 2 (76%). Moreover, case 2 retains high P for most of the threshold span. This supports the previous result indicating case 2 as the best choice.

(4) When it comes to AUC, case 2 outperformed all other cases for most of the threshold span. We consider this as a final support for case 2 obtaining the best performance. The results were represented in Fig. 5.

Table 4 BEP results for cases 1-6

case 1	case 2	case 3	case 4	case 5	case 6
0.56	0.59	0.55	0.52	0.53	0.54

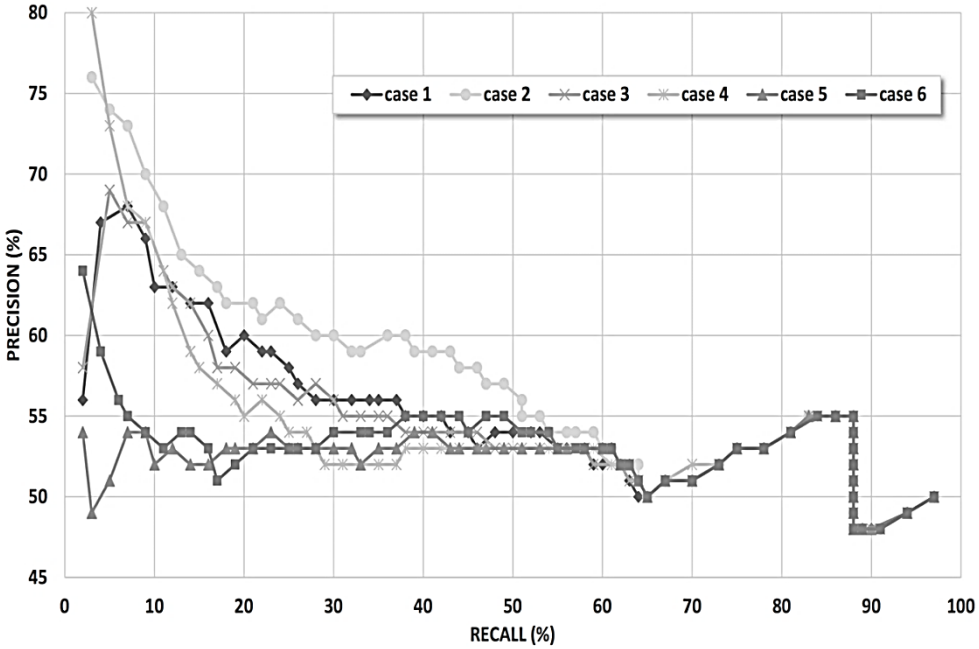

Fig. 5 Results of all compared cases

6. Discussion

Finally, we verified whether simply adding words to Nitta's method improves its performance. After analyzing all evaluation metrics, the cases' performance was in order: case 2 (12 words), case 1 (7w), case 3 (16w), case 4 (21w), case 6 (9w), case 5 (5w). The Spearman's Rank Correlation Test calculated between the ranks of cases and numbers of seed words was $\rho = -0.314$, which indicates no correlation with minor bias towards negative correlation. This suggests the following conclusions. Simply adding words does not improve the performance. Minor negative correlation could also suggest the contrary - less seed words results in better results. However, since there was no statistical significance for the correlation test (2-sided p-value = 0.564), to prove this more experiments need to be conducted on larger datasets.

Moreover, the two best scores were obtained by the cases applying the proposed double-filtering method. Here, case1 applies the full filtering procedure (seed17). Case 2 applies single filtering with final set of harmful seed words extracted with the help of Nitta et al.'s original seed set instead of double-filtered word set. This suggests that the filtering improves the performance of Nitta's method in general. Moreover, shortening the filtering from double- to single-filtering positively influences not only the performance, but also efficiency of the method. Applying only single filtering instead of double-filtering saves much of the processing time needed to extract the final set of words.

We verified the statistical significance of results using McNemar Test [5], a standard test used for calculating statistical significance for paired nominal data. Table 5 shows the McNemar Test results.

The comparison of the results of case 5 and case 6 with the results of other cases, indicated that a significant difference was obtained for all cases except for the difference of case 3 and case 4 with case 5.

Table 5 McNemar Test results (for cases 5 and 6 as baselines)

	case1	case2	case3	case4	case5	case6
case5	189.00 ***	26.88 ***	0.83	0.30	-	16.98 ***
case6	145.00 ***	5.80 *	9.47 **	10.29 **	18.56 ***	-

*p≦0.5, **p≦0.1, ***p≦0.01

7. Conclusions

In this paper, we presented our re-evaluation of Nitta et al.'s method for cyberbullying detection in informal BBS websites. We found out that the method originally proposed in 2013 lost over 30% points over two years. The loss in performance could be caused by several factors, including tightening of usage policies of many Internet services. Next, we proposed a method for double filtering of seed words using a method by Ishizaka et al. and performed an experiment on multiple automatically constructed seed word sets. The experiment results clearly showed that simply adding more words to Nitta et al.'s method is not effective. On the other hand, the proposed filtering method was significantly effective, although shortening the filtering from double to single improved not only efficiency of the method, but also its performance, regaining much of the originally reported performance. Moreover, the proposed method not only improved Nitta et al.'s method for the present time, but it can be used to improve it in the future, if the results drop again due to fluctuations in Web information.

In the near future, we plan to repeat the experiment by Ishizaka et al. to eliminate any potential ambiguities in human judgments about the harmfulness of words. We also plan to apply the proposed PMI filtering method to other Web-mining based methods and verify its potential and limitations in improving such methods in general.

References

[1] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, and K. Araki, "Detecting cyberbullying entries on informal school websites based on category relevance maximization," Proc. of the 6th International Joint Conference on Natural Language Processing (IJCNLP'13), Oct. 2013, pp. 579-586.

[2] A. Kilgarriff, "Googleology is bad science," Computational Linguistics, vol. 33, no. 1, pp. 147-151, 2007.

[3] T. Ishizaka and K. Yamamoto, "Automatic detection of sentences representing slandering on the Web," Proc. of the 17th Annual Meeting of the Association for Natural Language Processing, 2011, pp. 131-134. (In Japanese)

[4] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Proc. of ACL-02, Jul. 2002, pp. 417-424.

[5] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153-157, 1947.