A Study of Unsounded Code Strings at the End of Online Messages of a Q&A Site and a Micro Blog

Kunihiro Nakajima, Yasuhiko Watanabe, Subaru Nakayama, Kenji Umemoto, Ryo Nishimura, and Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
Email: t13m071@mail.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp, t090433@mail.ryukoku.ac.jp,
t11m074@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp, okada@rins.ryukoku.ac.jp

***Abstract*—In this study, we compare answers in a Q&A site with messages in a micro blog and discuss how we use unsounded code strings at the end of online messages. We first show that unsounded code strings at the end of answers in a Q&A site are used for not only smooth communication but an other purpose, minimum length limit avoidance. Next, we show that the length of unsounded code strings at the end of answers in a Q&A site, which are used for smooth communication, have a similar distribution pattern to those at the end of messages in a micro blog. On the other hand, the length of unsounded code strings used for minimum length limit avoidance have a different distribution pattern. Furthermore, we compare frequently used unsounded code strings at the end of answers in a Q&A site with those at the end of messages in a micro blog. Finally, we show our results are useful to analyze users' messages in online communities. In this study, we used the data of Yahoo! chiebukuro, a widely-used Japanese Q&A site, and Twitter for observation and examination.**

***Keywords*—unsounded code string, micro blog, Twitter, Q&A site, Yahoo! chiebukuro.**

I. INTRODUCTION

We often find consecutive unsounded marks and characters are used at the end of online messages, such as mails, chattings, and questions and answers in Q&A sites. As a result, it is important to investigate how these expressions were used.

(exp 1) *sound recorder demo aru teido ha dekiru kedo, yappari Sound Engine ga osusume kana...* (You may be able to do a lot by using sound recorders, however, the one I would like to recommend is Sound Engine...)

(exp 1) is an answer submitted to a Japanese Q&A site, Yahoo! chiebukuro. In this case, periods are used consecutively at the end of it. It is probable that the answerer of (exp 1) used the three consecutive periods for expressing his/her opinion gently, in other words, for smooth communication. In this study, we define unsounded marks and characters as *unsounded codes*. Furthermore, we define three or more consecutive unsounded codes as *unsounded code strings*. For example, in Yahoo! chiebukuro, 25 % of answers have unsounded code strings, in other words, three or more consecutive unsounded codes at the end of them. Although unsounded code strings are popular, there are few studies on them. As a result, we investigated how we use unsounded code strings at the end of online messages [1]. In the report, we compared answers in a Q&A site with messages in a micro blog and discussed how we use unsounded code strings at the end of online messages. We used the data of Yahoo! chiebukuro [2], a widely-used Japanese Q&A site, and Twitter [3] for observation and examination. In this study, we review our previous report and show the new results of our study. Especially, we compare frequently used unsounded code strings at the end of answers in a Q&A site with those at the end of messages in a micro blog. The results of this study will give us a chance to understand not only the usage of unsounded code strings in online messages but the purposes and behaviors of users in online communities. Especially, the results can be useful to analyze the impacts of communication constraints on users' messages and communications. In this paper, we show our results are useful to analyze the impacts of the minimum length limit in Yahoo! chiebukuro on users' messages and communications.

The rest of this paper is organized as follows: In Section II, we surveys the related works. In Section III, we describe how unsounded code strings are used at the end of answers in a Q&A site. On the other hand, in Section IV, we describe how unsounded code strings are used at the end of messages in a micro blog. Furthermore, we compare unsounded code strings at the end of answers in a Q&A site with those at the end of messages in a micro blog. In Section V, we show our results are useful to analyze users' messages in online communities. Finally, in Section VI, we present our conclusions.

II. RELATED WORKS

Yamamoto pointed out that the number of users in communication media for exchanging short text messages has been increasing rapidly [4]. One good example is Twitter. Twitter has succeeded in winning the hearts and minds of many users. The reason is to limiting the message length to 140 characters. By limiting the message length to 140 characters, Twitter has succeeded in encouraging users to submit many messages quickly and enhancing their communications. As a result, in order to develop new communication media technology, we

TABLE I. THE NUMBERS OF QUESTIONERS AND ANSWERERS IN YAHOO! CHIEBUKURO (FROM APRIL/2004 TO OCTOBER/2005).

	number of questioners	number of answerers
the data of Yahoo! chiebukuro	165,064	183,242

TABLE II. THE NUMBERS OF QUESTIONS AND ANSWERS IN YAHOO! CHIEBUKURO (FROM APRIL/2004 TO OCTOBER/2005).

	number of questions	number of answers
the data of Yahoo! chiebukuro	3,116,009	13,477,785

should investigate short text messages. For example, online messages in Twitter consist of

- strings for reference (URL, username, hashtag, etc.),
- utterable words, and
- unsounded code strings.

Unsouded code strings are used frequently in short text messages. However, there are few studies on unsounded code strings, except emoticons. As a result, in order to develop new communication media technology, it is important to investigate unsounded code strings.

Emoticons, sometimes called face marks, are a kind of unsounded code strings. First emoticon, smiley face ";-)", was proposed by Scott Fahlman in September 1982 [5]. After his proposal, many emoticons have been used widely in online messages, such as email, chat, and newsgroup posts [6]. As a result, a large number of studies have been made on emoticons.

Many researchers in computational linguistics proposed methods of extracting and classifying emoticons in online messages. Inoue et al. analyzed 1,000 sentences in email messages and developed a system which extracted emotional expressions, especially emoticons, embedded in email messages [7]. Nakamura et al. proposed a method of learning emoticons for a natural language dialogue system from chat dialogue data in the Internet [8]. Tanaka et al. proposed methods for extracting emoticons in text and classifying them into some emotional categories [9]. Bedrick et al. proposed robust emoticon detection method based on weighted context-free grammars [10]. Hogenboom et al. showed that sentiment classification accuracy was improved by using manually created emoticon sentiment lexicon [11].

On the other hand, many researchers in social science analyzed how we use emoticons in online messages. Witmer and Katzman reported that women use more graphic accents (emoticons) than men do in their computer-mediated communication (CMC) [12]. Walther and D'Addario showed that emoticons' contributions were outweighed by verbal content [13]. Derks et al. reported emoticons are useful in strengthening the intensity of a verbal message [14]. Byron and Baldridge reported readers were likely to rate sender's emails more likeable if they used emoticons [15]. Harada discussed how Japanese speakers use emoticons for promoting communication smoothly from the viewpoint of politeness [16]. Kato et al. analyzed positive and negative emoticons and reported that negative emoticons are misinterpreted more frequently than positive ones [17]. Furthermore, Kato et al. reported that emoticons are used more frequently between close friends than ordinary acquaintances [18].

We think emoticons are a kind of unsounded code strings, however, there are few studies on other kinds of unsounded code strings. As a result, we should investigate not only emoticons but other kinds of unsounded code strings. The results of this study are useful to understand the usage of unsounded code strings in online messages. Furthermore, the results will give us a chance to understand the purposes and behaviors of users in online communities.

III. UNSOUNDED CODE STRINGS AT THE END OF ANSWERS IN A Q&A SITE

In this section, we discuss unsounded code strings at the end of answers submitted to a Q&A site.

Before we define a unsounded code string, we explain the data of Yahoo! chiebukuro, which we used for investigating unsounded code strings in a Q&A site. Yahoo! chiebukuro is a Japanese version of Yahoo! answers and one of the most popular Q&A sites in Japan. In Yahoo! chiebukuro, 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 *best answer*. The data of Yahoo! chiebukuro was published by Yahoo! JAPAN via National Institute of Informatics in 2007 [19]. This data consists of about 3.11 million questions and 13.47 million answers which were posted on Yahoo! chiebukuro from April/2004 to October/2005. In the data, each question has at least one answer because questions with no answers were removed. 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 questioners and answerers in the data of Yahoo! chiebukuro. Table II shows the numbers of their questions and answers in the data of Yahoo! chiebukuro.

Next, we define an unsounded code and unsounded code strings. In this study, we define that an unsounded code string is three or more consecutive unsounded codes. In this study, unsounded codes are limited to the following marks and characters:

- punctuation marks,
- Greek characters,
- Cyrillic characters, and
- ruled lines.

These marks and characters are generally unsounded when they are used at the end of Japanese sentences. We observed unsounded code strings at the end of answers submitted to Yahoo! chiebukuro, and found they were used for

1) smooth communications

(exp 2) *koko ni kaki shirushita bunmen wo sonomama kanojyo ni misete ageru koto wo osusume shimasu. futari no aida ni shinrai kankei ga kizukete iru nara kitto daijyobu!!!* (You had better show what you described here to your girl friend with no change

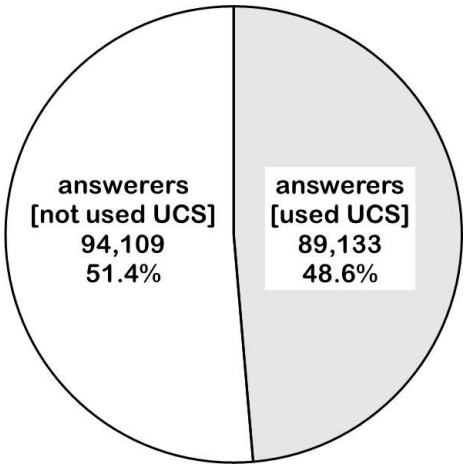

Fig. 1. The proportion and number of answerers who used unsounded code strings at the end of their answers (from April/2004 to October/2005). UCS means an unsounded code string.

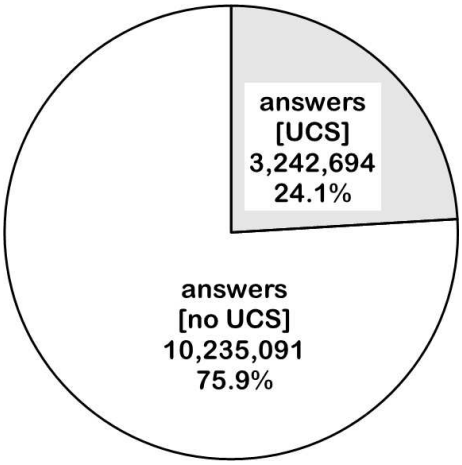

Fig. 2. The proportion and number of answers that have unsounded code strings at the end of them (from April/2004 to October/2005).

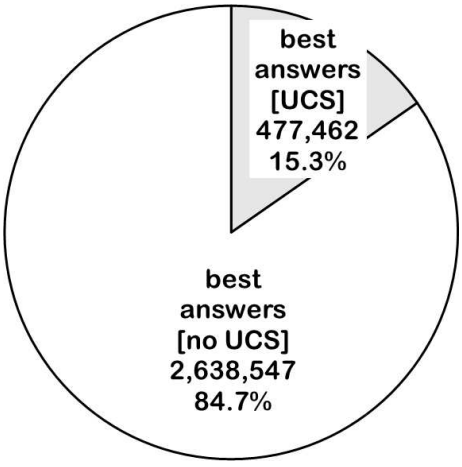

Fig. 3. The proportion and number of best answers that have unsounded code strings at the end of them (from April/2004 to October/2005).

at all. If you have a trust relationship with her, you don't worry!!!)

2) minimum length limit avoidance

(exp 3) *alumi foiru ni tsutsun de hi no naka ni pon!!!!!!!!!!!!!* (Wrap aluminum foil around and pop it into a fire!!!!!!!!!!!!!)

The minimum length limit was introduced into Yahoo! chiebukuro in May/2004. Due to this limit, users in Yahoo! chiebukuro are prohibited from submitting answers less than 25 multibyte characters long. Makoto Okamoto [20], a former producer of Yahoo! chiebukuro, said that this rule was introduced for avoiding less informative answer submissions. In this rule, one single byte character is counted as 0.5 multibyte character. In order to avoid this limit, the answerer of (exp 3) used 13 "!" at the end of his/her answer. We may note that, in case of Japanese texts, the length of words and sentences are generally counted by multibyte characters. In this study, single byte characters are counted as 0.5 multibyte characters. We count characters in the data of Yahoo! chiebukuro by using programming language Perl (version 5.14.2) [21] on Ubuntu linux (version 12.04) [22].

As shown in Table I, there are 183,242 users each of whom submitted at least one answer to Yahoo! chiebukuro. Figure 1 shows the proportion and number of users who used unsounded code strings at the end of their answers. As shown in Table II, 13,477,785 answers were submitted to Yahoo! chiebukuro, and 3,116,009 of them were selected as best answers. Figure 2 shows the proportion and number of answers that have unsounded code strings at the end of them. On the other hand, Figure 3 shows the proportion and number of best answers that have unsounded code strings at the end of them.

Figure 4 shows the cumulative relative frequency distribution of

- the length of all the answers,
- the length of answers that have unsounded code strings at the end of them, and
- the length of unsounded code strings.

As shown in Figure 4, the median of the length of unsounded code strings at the end of answers is 10 multibyte characters. This value is more than twice the length of unsounded code strings at the end of (exp 1) and (exp 2). We think that it is too long for smooth communication. As a result, we investigate the association between the length of

- unsounded code string at the end of answers and
- the other part of them.

The result is shown in Figure 5. In Figure 5, the heatmap shows the association between the length of unsounded code string at the end of answers and the other part of the answers. In the heatmap, darker color denotes more frequent data element. The heatmap shows long unsounded code strings at the end of answers are mainly used when the other part of the answers are less than 25 multibyte characters long. Furthermore, unsounded code strings at the end of the answers

TABLE III. THE NUMBER OF ANSWERERS, ANSWERS, AND BEST ANSWERS IN CASE OF ANSWERS THE LENGTH OF WHICH, EXCLUDING UNSOUNDED CODE STRINGS AT THE END OF THEM, WERE (1) LESS THAN 25 MULTIBYTE CHARACTERS AND (2) 25 MULTIBYTE CHARACTERS OR LONGER.

length of answers (excluding unsounded code strings at the end of them)	number of answerers	number of answers	number of best answers	best answer ratio
less than 25 multibyte characters	52,998	1,745,797	191,791	11.0
25 multibyte characters or longer	77,299	1,496,897	285,671	19.1

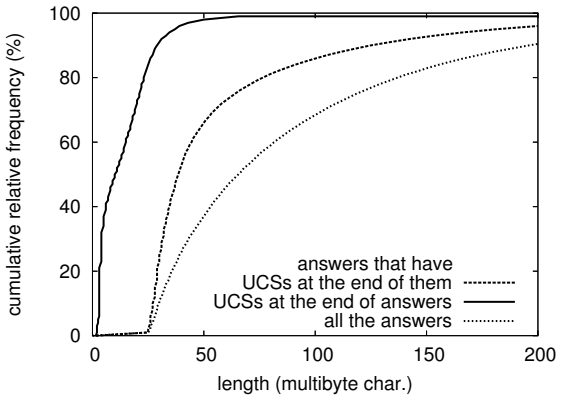

Fig. 4. The cumulative relative frequency distribution of the length of (1) all the answers, (2) answers that have unsounded code strings at the end of them, and (3) unsounded code strings at the end of them. UCS means an unsounded code string.

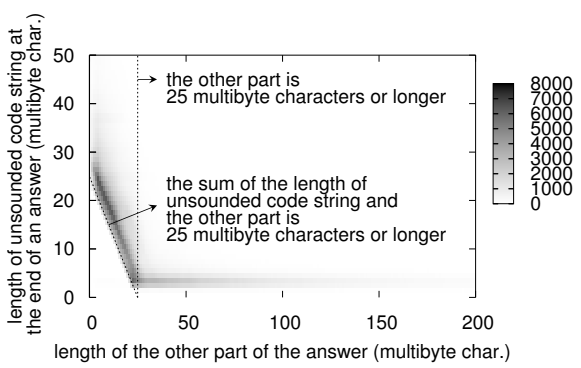

Fig. 5. The heatmap which shows the association between the length of unsounded code string at the end of answers and the other part of the answers.

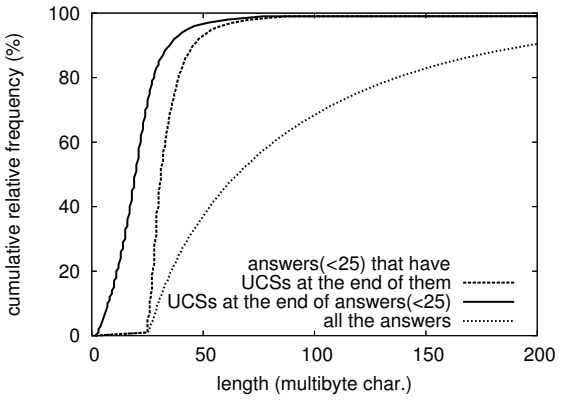

Fig. 6. The cumulative relative frequency distribution of the length of (1) all the answers, (2) answers that have unsounded code strings at the end of them and are less than 25 multibyte characters long (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

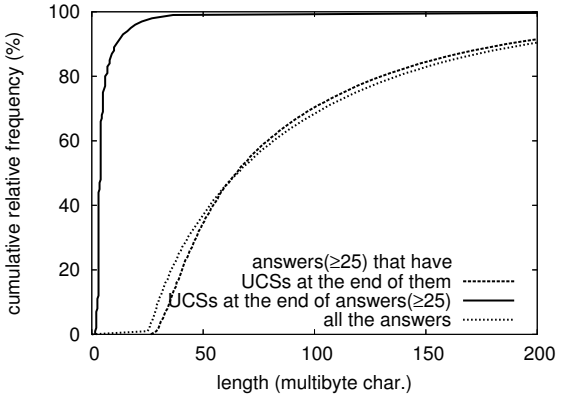

Fig. 7. The cumulative relative frequency distribution of the length of (1) all the answers, (2) answers that have unsounded code strings at the end of them and are 25 multibyte characters or longer (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

come in a variety of lengths, however, the sum of the length of unsounded code string at the end and the other part of them, in other words, the length of the answers are frequently 25–30 multibyte characters long. On the other hand, when the other part of answers are more than 25 multibyte characters long, unsounded code strings at the end of the answers are mainly 3–5 multibyte characters long, and the answers come in a variety of lengths. It may be said that the usage of unsounded code strings at the end of answers differs greatly depending on whether the other part of the answers are less than 25 multibyte characters long. As a result, we divided answers that have unsounded code strings at the end of them into

- answers the length of which are less than 25 multibyte characters (excluding unsounded code strings at the end of them)
- answers the length of which are 25 multibyte characters or longer (excluding unsounded code strings at the end of them)

and investigated them in the following points:

- the number of answerers, answers, and best answers

TABLE IV. THE TOP 40 MOST FREQUENTLY USED UNSOUNDED CODE STRINGS AT THE END OF ANSWERS THE LENGTH OF WHICH ARE LESS THAN 25 MULTIBYTE CHARACTERS (EXCLUDING UNSOUNDED CODE STRINGS AT THE END OF THEM)

unsounded code string	frequency
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	22,091
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	20,936
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	20,655
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	20,467
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	20,257
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	20,147
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	19,989
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	19,624
｡｡｡｡｡｡｡｡｡｡｡｡｡	18,722
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	18,552
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	17,718
｡｡｡｡｡｡｡｡｡｡｡	17,491
｡｡｡｡｡｡｡｡｡｡｡｡	17,475
｡｡｡｡｡｡｡｡｡｡	16,681
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	16,173
｡｡｡｡｡｡｡｡｡	15,299
｡｡｡｡｡｡｡｡	13,954
･･･････････････････	13,445
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	13,411
｡｡｡｡｡	13,163
･･･････････････････	13,134
････････････････････	13,122
｡｡｡｡｡｡｡	13,041
｡｡｡｡｡｡	13,003
･･･････････････････	12,901
･･･････････････････	12,867
･･･････････････････	12,148
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	12,113
･･･････････････････	12,099
･･･････････････････	12,051
｡｡｡	11,460
･･･････････････････	11,322
･･･････････････････	10,787
･･･････････････････	10,251
｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡｡	10,193
｡｡｡｡	10,133
･･･････････････	9,818
･･･････････････	9,423
･･･････････････････	9,284
･･･････････････	8,573

TABLE V. THE TOP 40 MOST FREQUENTLY USED UNSOUNDED CODE STRINGS AT THE END OF ANSWERS THE LENGTH OF WHICH ARE 25 MULTIBYTE CHARACTERS OR LONGER (EXCLUDING UNSOUNDED CODE STRINGS AT THE END OF THEM)

unsounded code string	frequency
･･･	205,483
･･･ ｡	137,534
｡｡｡	119,480
｡ ␣␣	65,237
････	42,212
!!!	33,206
･････	25,626
｡｡｡｡	25,306
???	21,107
･･ ｡	19,694
｡｡｡｡｡	18,012
････ ｡	17,387
･･･････	13,995
｡ ␣␣␣	12,453
｡｡｡｡｡｡	9,295
･････ ｡	9,094
!!!!	8,614
^^;	8,391
･･･････	8,268
･･･ ␣	7,905
･･･?	7,593
｡｡｡｡｡｡｡	6,943
``	6,764
････････	6,322
!!!!!	6,277
｡｡｡｡｡｡｡｡	6,082
? ␣␣	5,762
･･･ ｡ ␣	5,725
␣*･*･*･*･*	5649
｡｡｡｡｡｡｡｡｡	5453
･･･････････	5439
｡｡｡｡｡｡｡｡｡｡	4990
･･･)	4875
(^^)	4767
･･････････	4715
｡ ␣␣␣␣	4560
!! ␣	4537
????	4482
･･･････ ｡	4473
｡｡｡｡｡｡｡｡｡｡｡	4459

␣ means a single byte space.

(Table III),

- the length of answers and unsounded code strings at the end of them (Figure 6 and Figure 7), and
- frequently used unsounded code strings at the end of answers (Table IV and Table V).

First, we discuss answers the length of which are less than 25 multibyte characters (excluding unsounded code strings at the end of them). In case of these answers, unsounded code strings at the end of them were used for avoiding the minimum length limit. This limit is a special problem in Yahoo! chiebukuro, not introduced into Twitter. As a result, we do not compare unsounded code strings for avoiding the minimum length limit with those used at the end of online messages in Twitter.

Next, we discuss answers the length of which are 25 multibyte characters or longer (excluding unsounded code strings at the end of them). In case of these answers, unsounded code strings at the end of them were used for smooth communication, not for minimum length limit avoidance. As shown in

Figure 7, the length of these answers (excluding unsounded code strings at the end of them) have a distribution similar to those of all the answers submitted to Yahoo! chiebukuro. As a result, it may be said that, when the length of answers are 25 multibyte characters or longer (excluding unsounded code strings at the end of them), the length of these answers are less affected by whether unsounded code strings are used at the end of them. We compare these unsounded code strings with those used at the end of online messages in Twitter.

Finally, we discuss frequently used unsounded code strings at the end of answers in Yahoo! chiebukuro. Table IV shows the top 40 most frequently used unsounded code strings at the end of answers the length of which are less than 25 multibyte characters (excluding unsounded code strings at the end of them). As shown in Table IV, all of the top 40 most frequently used unsounded code strings are classified into two types of consecutive multibyte characters, consecutive Japanese periods, such as "。。。", or consecutive multibyte bullets, such as "・・・". These consecutive multibyte characters occupy 43.2 % of all the unsounded code strings at the end of answers the length of which are less than 25 multibyte characters. On the other hand, Table V shows the top 40 most frequently used unsounded code strings at the end of answers the length of which are 25 multibyte characters or longer (excluding unsounded code strings at the end of them). As shown in Table V, 17 of the top 40 most frequently used unsounded code strings are also these consecutive multibyte characters. These consecutive multibyte characters occupy 37.2 % of all the unsounded code strings at the end of answers the length of which are 25 multibyte characters or longer. As a result, in both cases, unsounded code strings composed only of these consecutive multibyte characters are used frequently. However, the distribution of the length of them differ greatly. The unsounded code strings at the end of answers the length of which are less than 25 multibyte characters are mainly long. Only 4.5 % of them are "。。。", "。。。。", "・・・", or "・・・・". On the other hand, the unsounded code strings at the end of answers the length of which are 25 multibyte characters or longer are mainly short. 70.7 % of them are "。。。", "。。。。", "・・・", or "・・・・". There is another thing to note. Emoticons are not popular at the end of answers in Yahoo! chiebukuro. As shown in Table V, only two types of emoticons, ^^; and (^^), are ranked in the top 40. We think that it is partially due to the prohibition of ASCII and Shift JIS artworks. ASCII and Shift JIS artworks are often found in blog comment forms, Q&A sites, and web-based bulletin boards. Figure 8 shows an typical example of Shift JIS artwork submitted to a Japanese web-based bulletin board, 2channel [23]. However, operators in Yahoo! chiebukuro thought this kind of submissions, especially, ASCII and Shift JIS artworks, such as Figure 8, were not good for communications in Yahoo! chiebukuro, and then, they prohibited users from submitting them. We think many users in Yahoo! chiebukuro reduced the use of emoticons because they think that emoticons are classified into ASCII and Shift JIS artworks. We will discuss another reason why many users in Yahoo! chiebukuro reduced the use of emoticons in next section.

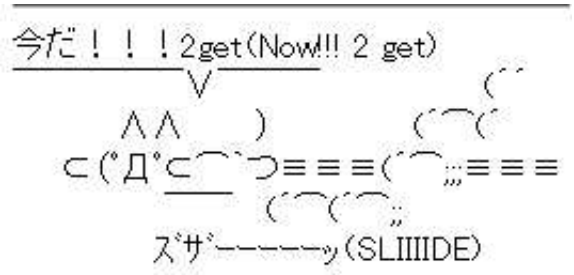

Fig. 8. An example of Shift JIS artwork to a Japanese web-based bulletin board, 2channel.

IV. UNSOUNDED CODE STRINGS AT THE END OF MESSAGES IN A MICRO BLOG

In order to compare with unsounded code strings at the end of answers in Yahoo! chiebukuro, we investigate unsounded code strings at the end of messages in Twitter. We obtained messages submitted to Twitter, in other words, tweets by using the streaming API. However, the streaming API allows us to obtain only 1% of all public streamed tweets because of API restriction. We used the streaming API and obtained 7,085,267 Japanese tweets in three weeks in November and December 2012. These tweets can be classified into three types:

- reply
  A reply to a particular user. It contains "@username" in the body of the tweet.
- retweet
  A retweet is a reply to a tweet that includes the original message.
- normal tweet
  A normal tweet is neither reply nor retweet.

Figure 9 shows the proportion and number of normal tweets, replies, and retweets in Twitter (from November/2012 to December/2012). From these tweets, we extracted 966,187 tweets that have unsounded code strings at the end of them. These 966,187 tweets are 13.6% of all the tweets. Figure 10 shows the proportion and number of normal tweets, replies, and retweets that have unsounded code strings at the end of them. As shown in Figure 9 and Figure 10, replies occupy 44.5% of tweets that have unsounded code strings at the end of them while replies occupy 35.7% of all the tweets. As a result, replies have unsounded code strings at the end of them more frequently than other kinds of tweets. It is because each reply is sent to a particular person. When we send a message to a particular person, we generally try to avoid unnecessary frictions with him/her. As a result, we use unsounded code strings at the end of our replies more frequently than other kinds of tweets.

Before we discuss unsounded code strings at the end of tweets, we remove retweets. It is because, messages in retweets are created not by submitters, but by other users. As a result, retweets are inadequate to investigate how we use unsounded code strings at the end of online messages. Figure 11 shows the cumulative relative frequency distribution of

- the length of all the tweets (excluding retweets),

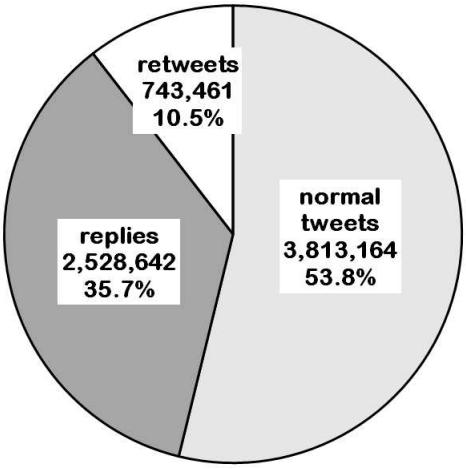

Fig. 9. The proportion and number of normal tweets, replies, and retweets in Twitter (from November/2012 to December/2012).

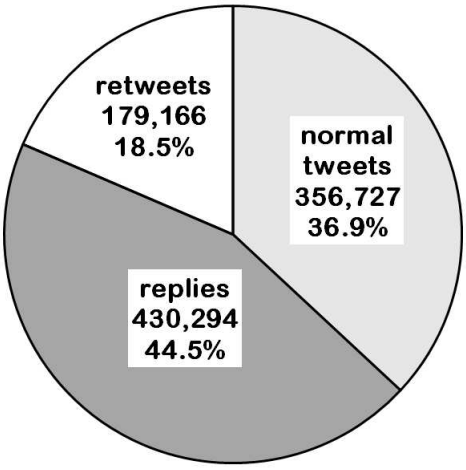

Fig. 10. The proportion and number of normal tweets, replies, and retweets in Twitter that have unsounded code strings at the end of them (from November/2012 to December/2012).

- the length of tweets (excluding retweets) that have unsounded code strings at the end of them, and
- the length of unsounded code strings at the end of tweets (excluding retweets).

In Figure 12, the heatmap shows the association between the length of unsounded code string at the end of tweets and the other part of the tweets. Figure 11 and Figure 12 show unsounded code strings at the end of the tweets are mainly 3–5 multibyte characters long, and the tweets come in a variety of lengths. The length of unsounded code strings at the end of tweets have a similar distribution pattern to those of answers in Yahoo! chiebukuro, which are 25 multibyte characters or longer (excluding unsounded code strings at the end of them). As a result, unsounded code strings at the end of online

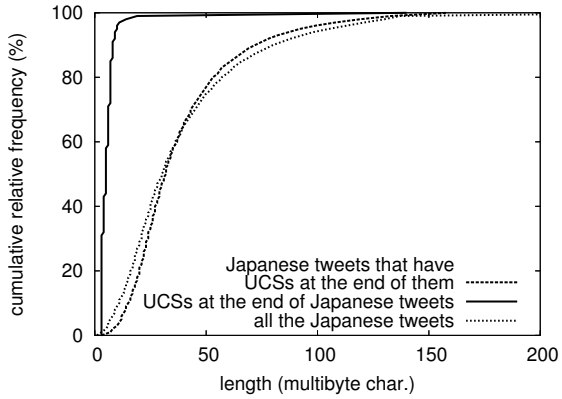

Fig. 11. The cumulative relative frequency distribution of the length of (1) all the Japanese tweets, (2) Japanese tweets that have unsounded code strings at the end of them, and (3) unsounded code strings at the end of them.

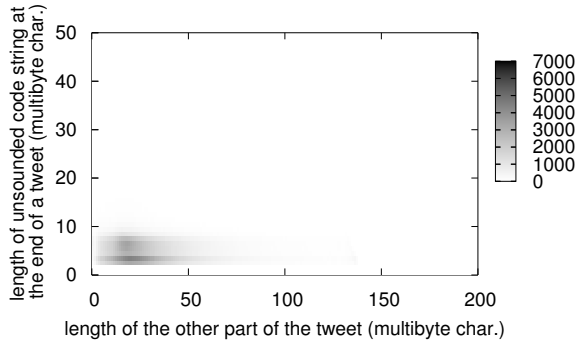

Fig. 12. The heatmap which shows the association between the length of unsounded code string at the end of Japanese tweets and the other part of the tweets.

messages are mainly 3–5 multibyte characters long when they are used for smooth communications with particular persons. Table VI shows the top 40 most frequently used unsounded code strings at the end of tweets (excluding retweets). As shown in Table VI, only one kind of consecutive Japanese periods, “。。。”, and two kinds of consecutive multibyte bullets, “・・・” and “・・・・”, are ranked in the top 40 most frequently used unsounded code strings at the end of tweets (excluding retweets). These consecutive multibyte characters occupy 5.6 % of all the unsounded code strings at the end of tweets (excluding retweets). As a result, these consecutive multibyte characters, such as “。。。” and “・・・”, are used less frequently at the end of tweets than at the end of answers in Yahoo! chiebukuro. On the other hand, emoticons are used more frequently at the end of tweets than at the end of answers in Yahoo! chiebukuro. 24 of the top 40 most frequently used unsounded code strings at the end of tweets (excluding retweets) are emoticons or parts of emoticons. One of the reasons why emoticons are used frequently at the end of tweets is that Twitter users often sent their tweets to

TABLE VI. THE TOP 40 MOST FREQUENTLY USED UNSOUNDED CODE STRINGS AT THE END OF TWEETS (EXCLUDING RETWEETS).

unsounded code string	frequency
^) ／	32518
•••	30710
!!!	26511
)))	13847
(*^^*)	11665
...	10603
!!!!	10307
ー !!	7396
(^^)	7096
°°°	7036
(´• ω •`)	6093
!!!!!	5962
///	5371
(￣▽￣)	5013
! ! !	4823
(^−^)	4281
^)/*	4128
(* ´ ω ` *)	4001
••• ｡	3644
(´; ω ;`)	3458
(´Д‘␣)	3429
(´• ω • ‘)	3361
(^−^)/	3333
!!!!!!	3218
‘*)	3200
(;_;)	3149
???	3123
••••	3028
^)/	3015
ーーー	2989
␣」 ∠)_	2943
^*)	2916
」 ∠)_	2905
(^_^;)	2886
ー !!!	2736
……｡	2695
……	2687
^_^	2680
``	2646
)))))))	2637

␣ means a single byte space.

familiar persons while answerers in Yahoo! chiebukuro almost always sent their answers to strangers. Kato et al. reported that emoticons are used more frequently between close friends than ordinary acquaintances [18]. As a result, we think that emoticons are used more frequently in replies than in normal tweets. It is because replies are sent to particular persons. On the other hand, normal tweets are sent to not only particular persons but general public.

Next, we discuss unsounded code strings at the end of normal tweets and replies, individually. Figure 13 shows the

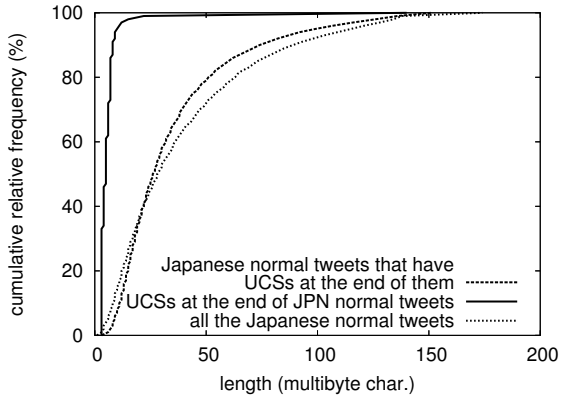

Fig. 13. The cumulative relative frequency distribution of the length of (1) all the Japanese normal tweets, (2) Japanese normal tweets that have unsounded code strings at the end of them (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

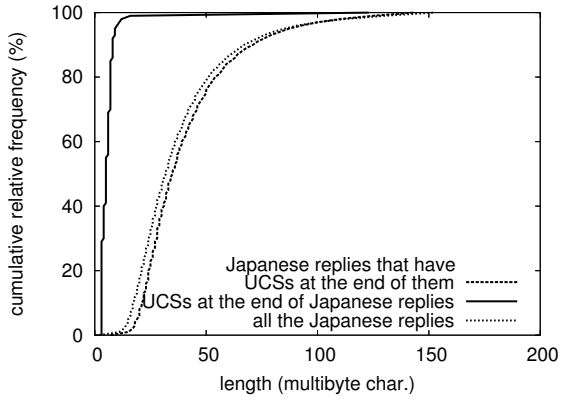

Fig. 14. The cumulative relative frequency distribution of the length of (1) all the Japanese replies, (2) Japanese replies that have unsounded code strings at the end of them (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

cumulative relative frequency distribution of

- the length of all the normal tweets,
- the length of normal tweets that have unsounded code strings at the end of them (excluding unsounded code strings at the end of them), and
- the length of unsounded code strings at the end of normal tweets.

Also, Figure 14 shows the cumulative relative frequency distribution of

- the length of all the replies,
- the length of replies that have unsounded code strings at the end of them (excluding unsounded code strings at the end of them), and
- the length of unsounded code strings at the end of replies.

As shown in Figure 14, there are few short replies, especially less than 5 multibyte long. It is because each reply includes "@username". Also, as shown in Figure 14, the length of

TABLE VII. THE TOP 40 MOST FREQUENTLY USED UNSOUNDED CODE STRINGS AT THE END OF NORMAL TWEETS.

unsounded code string	frequency
･･･	20699
!!!	13621
^) ／	11845
. . .	7188
!!!!	5691
。。。	4749
)))	4236
!!!!!	3463
￣ !!	3277
(´･ω･`)	3262
(*^^*)	2807
! ! !	2569
･･･。	2518
￣￣￣	2139
････	2055
␣」∠)_	1925
」∠)_	1922
(^^)	1904
!!!!!!	1890
……	1861
……。	1859
///	1817
(´･ω･‘)	1795
(´Д‘␣)	1736
(´;ω;`)	1661
‘*)	1647
``	1644
???	1557
(*´ω`*)	1490
(;_;)￣	1474
(　▽　)	1470
(^−^)	1411
￣ !!!	1370
! ! ! !	1303
(*_*)	1236
!!!!!!!	1192
^)/*	1165
(´Д`)	1155
･･･?	1102
)))))))	1065

␣ means a single byte space.

TABLE VIII. THE TOP 40 MOST FREQUENTLY USED UNSOUNDED CODE STRINGS AT THE END OF REPLIES.

unsounded code string	frequency
^) ／	20673
!!!	12890
･･･	10011
)))	9611
(*^^*)	8858
(^^)	5192
!!!!	4616
￣ !!	4119
///	3554
(￣▽￣)	3543
. . .	3415
^)/*	2963
(^−^)	2870
(´･ω･`)	2831
(^−^)/	2616
(*´ω`*)	2511
!!!!!	2499
^)/	2349
。。。	2287
! ! !	2254
^*)	2153
(^_^;)	1888
^ ^	1845
(￣;ω;`)	1797
(´Д‘␣)	1693
(;_;)	1675
(*^_^*)	1675
(*´▽`*)	1637
(␣^ω^␣)	1627
)))))))	1572
???	1566
(´･ω･‘)	1566
‘*)	1553
♪(␣´▽`)	1504
(*´∀`*)	1433
\(//▽//)\	1423
(^^)／	1378
￣ !!!_	1366
(￣￣￣)	1361
!!!!!!	1328

␣ means a single byte space.

replies that have unsounded code strings at the end of them have a similar distribution pattern to the length of all the replies. It may be said that the length of replies are less affected by whether unsounded code strings are used at the end of them. This result is similar to the result obtained when we investigated answers in Yahoo! chiebukuro. The length of answers in Yahoo! chiebukuro, which are 25 multibyte characters or longer (excluding unsounded code strings at the end of them), are less affected by whether unsounded code strings are used at the end of them. In both cases of Yahoo! chiebukuro and Twitter, unsounded code strings are used for smooth communication with particular persons. As a result, it may also be said that the length of online messages to particular persons are less affected by whether unsounded code strings for smooth communication are used at the end of them. On the other hand, as shown in Figure 13, the length of normal tweets that have unsounded code strings at the end of them have a slightly different distribution pattern to the length of all the normal tweets. It is because normal tweets were sent to not only particular persons but general public while replies

were sent to particular persons.

Table VII and Table VIII show the top 40 most frequently used unsounded code strings at the end of normal tweets and replies, respectively. As shown in Table VII, 20 kinds of emoticons or parts of emoticons are ranked in the top 40 most frequently used unsounded code strings at the end of normal tweets. These emoticons occupy 36.5 % of all the unsounded code strings ranked in the top 40 most frequently used unsounded code strings at the end of normal tweets. On the other hand, as shown in Table VIII, 29 kinds of emoticons or parts of emoticons are ranked in the top 40 most frequently used unsounded code strings at the end of replies. These emoticons occupy 67.3 % of all the unsounded code strings ranked in the top 40 most frequently used unsounded code strings at the end of replies. As a result, emoticons are used more frequently at the end of replies than at the end of normal tweets. In other words, emoticons are used more frequently at the end of tweets to particular persons than at the end of tweets to general public.

V. Discussions

In this section, we show our results are useful to analyze the impacts of communication constraints on users' messages and communications. We take the minimum length limit in Yahoo! chiebukuro for example.

After the minimum length limit was introduced, users in Yahoo! chiebukuro have been prohibited from submitting answers less than 25 multibyte characters long. However, our study showed 1,745,797 answers, that is, 13.0 % of all the answers in the data of Yahoo! chiebukuro, were less than 25 multibyte characters (excluding unsounded code strings at the end of them). These answers were submitted by 52,998 users, that is, 28.9 % of all the answerers in the data of Yahoo! chiebukuro. It shows that many users in Yahoo! chiebukuro wanted to submit short answers. Furthermore, our study showed unsounded code strings used for smooth communication are mainly 3–5 multibyte characters long. We therefore classify these 1,745,797 answers into two types:

- 1,642,866 answers the unsounded code strings at the end of which were more than 5 characters long, and
- 102,931 answers the unsounded code strings at the end of which were 3-5 characters long.

In the former case, most of the unsounded code strings were thought to be used for minimum length limit avoidance. These unsounded code strings were often unfit for the contents of answers and gave poor impressions to questioners. As a result, the best answer ratio of these 1,642,866 answers was 10.8 % while that of all the answers in the data of Yahoo! chiebukuro was 23.1 %. On the other hand, in the latter case, some of the unsounded code strings were thought to be used for smooth communication. However, the best answer ration of these 102,931 answers was 12.4 %. As a result, in both cases, the best answer ratios were lower than that of all the answers. This result shows short answers are often less informative than long answers, in other words, the minimum length limit is reasonable. However, we should not overlook the positive factor of short answers. Ohsawa et al. reported that short and less informative submissions sometimes promote constructive discussions in web-based bulletin boards [24]. As a result, it is probable that some short and less informative answers stimulate other answerers to submit their good answers and enhance communications in Yahoo! chiebukuro.

VI. Conclusion

In this study, we investigated unsounded code strings at the end of answers in Yahoo! chiebukuro and tweets in Twitter. Although unsounded code strings are popular, there were few studies on them.

In Twitter, unsounded code strings at the end of tweets are used for smooth communication. On the other hand, in Yahoo! chiebukuro, unsounded code strings at the end of answers are used for not only smooth communication but minimum length limit avoidance. The minimum length limit is a special problem in Yahoo! chiebukuro, not introduced into Twitter. We showed that the usage of unsounded code strings at the end of answers in Yahoo! chiebukuro differs greatly depending on whether answers are longer than the minimum length limit. When answers are longer than the minimum length limit, unsounded code strings at the end of them are used for smooth communication. In this case, the length of the unsounded code strings at the end of answers have a similar distribution pattern to the length of unsounded code strings at the end of tweets. Unsounded code strings at the end of the tweets in Twitter and answers in Yahoo! chiebukuro, which are longer than the minimum length limit, are mainly 3–5 multibyte characters long. In addition, we showed the length of replies in Twitter and answers in Yahoo! chiebukuro, which are larger than the minimum length limit, are less affected by whether unsounded code strings are used at the end of them. Furthermore, we showed frequently used unsounded code strings at the end of answers in Yahoo! chiebukuro and tweets in Twitter. We showed that emoticons were not used frequently at the end of answers in Yahoo! chiebukuro. On the other hand, they were used frequently at the end of tweets in Twitter, especially, replies. The difference is whether messages are submitted to familiar persons or not. In other words, emoticons are used more frequently at the end of messages which are sent to familiar persons than to strangers and general public. Finally, we took the minimum length limit in Yahoo! chiebukuro for example and showed our results could be useful to analyze the impacts of communication constraints on users' messages and communications.

In this study, we analyzed and compared unsounded code strings at the end of answers in Yahoo! chiebukuro and Japanese tweets in Twitter. However, it is not enough to obtain general knowledge about unsounded code strings. It is because we have found many unsounded code strings not only in Japanese tweets but also in other language tweets, for example, English, French, Spanish, Portuguese, and so on. We intend to study the usages of unsounded code strings in these languages and compare them with the usage of Japanese unsounded code strings. We think the results of our future work are useful to provide new multilingual communication services.

REFERENCES

[1] K. Nakajima, S. Nakayama, Y. Watanabe, K. Umemoto, R. Nishimura, and Y. Okada, "An analysis of unsounded code strings in online messages of a q&a site and a micro blog," in *Proc. INTERNET 2013*, 2013, pp. 32–37.

[2] *Yahoo! chiebukuro*, Yahoo! JAPAN, 2004. [Online]. Available: http://chiebukuro.yahoo.co.jp/ [retrieved: December, 2014]

[3] *Twitter*, Twitter, Inc., 2006. [Online]. Available: https://twitter.com/ [retrieved: December, 2014]

[4] Y. Yamamoto, "Next-generation communications media, considering the evolution," in *Technical Report of IEICE on Human Communication Science (HCS)*, vol. 109, no. 457, 2010, pp. 19–20.

[5] S. Fahlman, *SMILEY:31 YEARS OLD AND NEVER LOOKED HAPPIER!*, Carnegie Mellon School of Computer Science, 2012. [Online]. Available: http://www.cs.cmu.edu/smiley/ [retrieved: December, 2014]

[6] H. Nojima, "(smily face) as a mean for emotional communication in networks," in *Proc. IPSJ summer programming symposium*, 1989, pp. 41–48.

[7] M. Inoue, M. Fujimaki, and S. Ishizaki, "System for analyzing emotional expression in e-mail text: collection, classification, and analysis of emotional expressions," in *Technical Report of IEICE on Thought and Language (TL)*, vol. 96, no. 608, 1997, pp. 1–8.

[8] J. Nakamura, T. Ikeda, N. Inui, and Y. Kotani, "Learning face marks for natural language dialogue systems," in *Proc. 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 2003, pp. 180–185.

[9] Y. Tanaka, H. Takamura, and M. Okumura, "Extraction and classification of facemarks," in *Proceedings of the 10th international conference on Intelligent user interfaces*, 2005, pp. 28–34.

[10] S. Bedrick, R. Beckley, B. Roark, and R. Sproat, "Robust kaomoji detection in Twitter," in *Proceedings of the Second Workshop on Language in Social Media*, 2012, pp. 56–64.

[11] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting emoticons in sentiment analysis," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 703–710.

[12] D. F. Witmer and S. L. Katzman, "On-line smiles: Does gender make a difference in the use of graphic accents?" *Journal of Computer-Mediated Communication*, vol. 2, no. 4, 1997. [Online]. Available: http://jcmc.indiana.edu/vol2/issue4/witmer1.html [retrieved: December, 2014]

[13] J. B. Walther and K. P. D'Addario, "The impacts of emoticons on message interpretation in computer-mediated communication," *Social Science Computer Review*, vol. 19, no. 3, pp. 324–347, 2001.

[14] D. Derks, A. E. R. Bos, and J. von Grumbkow, "Emoticons and online message interpretation," *Social Science Computer Review*, vol. 26, no. 3, pp. 379–388, 2008.

[15] K. Byron and D. C. Baldridge, "Email recipients' impressions of senders likeability," *Journal of Business Communication*, vol. 44, pp. 137–160, 2007.

[16] T. Harada, "The role of "face marks" in promoting smooth communication and expressing consideration and politeness in japanese," *the journal of the Institute for Language and Culture*, vol. 8, pp. 205–224, 2004.

[17] S. Kato, Y. Kato, M. Kobayashi, and M. Yanagisawa, "Analysis of the kinds of emotions interpreted from the emoticons used in e-mail," *the journal of Japan Society of Educational Information*, vol. 22, no. 4, pp. 31–39, 2007.

[18] S. Kato, Y. Kato, Y. Shimamine, and M. Yanagisawa, "Analysis of functions of emoticons in e-mail communication by mobile phone: Investigation of effects of degrees of intimacy with partners," *the journal of Japan Society of Educational Information*, vol. 24, no. 2, pp. 47–55, 2008.

[19] *The data of Yahoo! chiebukuro*, National Institute of Informatics, 2007. [Online]. Available: http://research.nii.ac.jp/tdc/chiebukuro.html [retrieved: December, 2014]

[20] *Makoto Okamoto*, ACADEMIC RESOURCE GUIDE(ARG), 2009. [Online]. Available: http://www.arg.ne.jp/node/7000 [retrieved: December, 2014]

[21] *The Perl Programming Language*, Perl.org, 2002. [Online]. Available: http://www.perl.org/ [retrieved: December, 2014]

[22] *Ubuntu Japanese Team*, Ubuntu Japanese Team, 2005. [Online]. Available: http://www.ubuntulinux.jp/japaneseteam/ [retrieved: December, 2014]

[23] *2channel*, PACKET MONSTER INC., 1999. [Online]. Available: http://www.2ch.net/ [retrieved: January, 2014]

[24] Y. Ohsawa, N. Matsumura, and Y. NAKAMURA, "Flaming promotes argumentation : A case study on of "2channel"," in *Technical Report of IEICE on Internet Architecture (IA)*, vol. 102, no. 143, 2002, pp. 55–60.