An Analysis of Unsounded Code Strings in Online Messages of a Q&A Site and a Micro Blog

Kunihiro Nakajima, Subaru Nakayama, Yasuhiko Watanabe, Kenji Umemoto, Ryo Nishimura, Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
Email: {t13m071, t090433}@mail.ryukoku.ac.jp, watanabe@rins.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 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 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. 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 because the answerer of (exp 1) is thought to use 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, in this study, we investigate how we use unsounded code strings at the end of online messages. Especially, 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 used the data of Yahoo! chiebukuro [1], a widely-used Japanese Q&A site, and twitter for observation and examination. The results of this study will give us a chance to understand the usage of unsounded code strings, and the purposes and behaviors of users in online communities. Especially, the results could be useful to predict and analyze the impacts of communication constraints 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 describes how unsounded code strings used at the end of answers in a Q&A site. On the other hand, in Section IV, we describes how unsounded code strings used at the end of messages in a micro blog. Finally, in Section V, we present our conclusions.

II. RELATED WORKS

Emoticons, sometimes called face marks, are a kind of unsounded code strings. First emoticon, smiley face ";-)", was proposed by Scott Fahlman in September 1982 [2]. After his proposal, many emoticons have been used widely in online messages, such as email, chat, and newsgroup posts [3]. 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 1000 sentences in email messages and developed a system which extracted emotional expressions, especially emoticons, embedded in email messages [4]. Nakamura et al. proposed a method of learning emoticons for a natural language dialogue system from chat dialogue data in the Internet [5]. Tanaka et al. proposed methods for extracting emoticons in text and classifying them into some emotional categories [6]. Bedrick et al. proposed robust emoticon detection method based on weighted context-free grammars [7]. Hogenboom et al. showed that sentiment classification accuracy was improved by using manually created emoticon sentiment lexicon [8].

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) [9]. Walther and D'Addario showed that emoticons' contributions were outweighed by verbal content [10]. Derks et al. reported emoticons are useful in strengthening the intensity of a verbal message [11]. Byron and Baldridge reported readers were likely to rate sender's emails more likeable if they used emoticons [12]. Harada discussed how Japanese speakers use emoticons for promoting communication smoothly from the viewpoint of politeness

 ISBN: 978-1-61208-285-1

TABLE I. THE NUMBERS OF USERS AND THEIR MESSAGES (QUESTIONS AND ANSWERS) SUBMITTED TO YAHOO! CHIEBUKURO (FROM APRIL/2004 TO OCTOBER/2005).

	number of questioners	number of questions	number of answerers	number of answers
the data of Yahoo! chiebukuro	165,064	3,116,009	183,242	13,477,785

TABLE II. THE NUMBER OF ANSWERERS, ANSWERS, AND BEST ANSWERS IN CASE OF (1) ALL THE ANSWERS IN YAHOO! CHIEBUKURO AND (2) ANSWERS WHICH HAVE UNSOUNDED CODE STRINGS AT THE END OF THEM.

	number of answerers	number of answers	number of best answers
all the answers	183,242	13,477,785	3,116,009
answers which have unsounded code strings at the end of them	89,133	3,242,694	477,462

[13]. Kato et al. analyzed positive and negative emoticons and reported that negative emoticons are misinterpreted more frequently than positive ones [14]. Furthermore, Kato et al. reported that emoticons are used more frequently between close friends than ordinary acquaintances [15].

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 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 [16]. 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 users and 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 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. 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.

Table II shows the number of answerers, answers, and best answers, in case of all the answers submitted to Yahoo! chiebukuro, and answers which have unsounded code strings at the end of them.

Figure 1 shows the cumulative relative frequency distribution of

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

As shown in Figure 1, the median of the length of unsounded code strings at the end of answers is 10 multibyte characters.

 ISBN: 978-1-61208-285-1

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
less than 25 multibyte characters	52,998	1,745,797	191,791
25 multibyte characters or longer	77,299	1,496,897	285,671

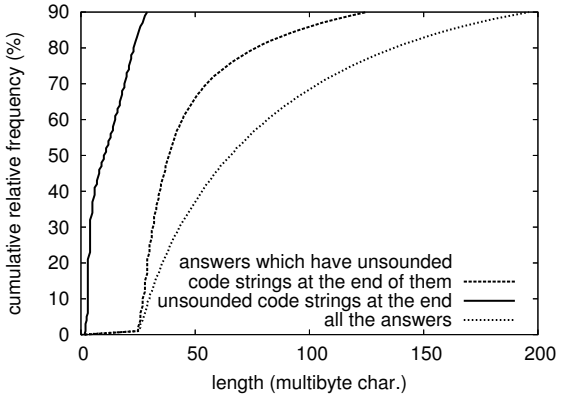

Fig. 1. The cumulative relative frequency distribution of the length of (1) all the answers, (2) answers which have unsounded code strings at the end of them, and (3) unsounded code strings at the end of them.

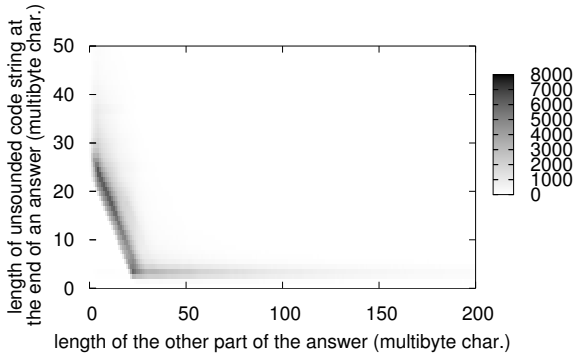

Fig. 2. The heatmap which shows the association between the length of unsounded code string at the end of answers and the other part of the answers.

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 2. In Figure 2, 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 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

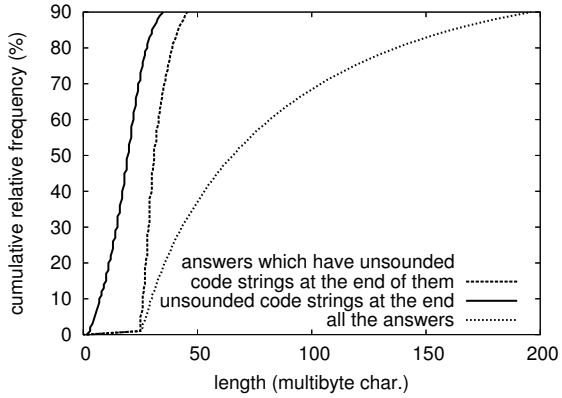

Fig. 3. The cumulative relative frequency distribution of the length of (1) all the answers, (2) answers which are less than 25 multibyte characters long (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

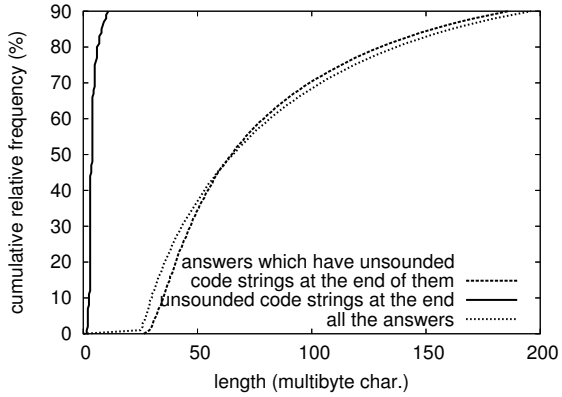

Fig. 4. The cumulative relative frequency distribution of the length of (1) all the answers, (2) answers which are 25 multibyte characters or longer (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

than 25 multibyte characters long, unsounded code strings at the end of the answers are mainly 3–4 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 which 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

TABLE IV. THE NUMBERS OF NORMAL TWEETS, REPLIES, AND RETWEETS IN TWITTER (FROM NOVEMBER/2012 TO DECEMBER/2012).

type of tweets	number of tweets	
normal	3,823,066	(53.97%)
reply	2,517,781	(35.54%)
retweet	743,336	(10.49%)
total	7,084,183	(100.00%)

TABLE V. THE NUMBERS OF NORMAL TWEETS, REPLIES, AND RETWEETS WHICH HAVE UNSOUNDED CODE STRINGS AT THE END OF THEM (FROM NOVEMBER/2012 TO DECEMBER/2012).

type of tweets	number of tweets	
normal	439,639	(38.15%)
reply	527,257	(45.75%)
retweet	185,547	(16.10%)
total	1,152,443	(100.00%)

the end of them)

and investigated them in the following points:

- the number of answerers, answers, and best answers (Table III),
- the length of answers and unsounded code strings at the end (Figure 3 and Figure 4),

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 in online messages of 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 4, 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 in online messages of twitter.

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 7,084,183 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.

Table IV shows the number of normal tweets, replies, and retweets. From these tweets, we extracted 1,152,443 tweets which have unsounded code strings at the end of them. These 1,152,443 tweets are 16.27% of all the tweets. Table V shows the number of normal tweets, replies, and retweets which have unsounded code strings at the end of them. As shown in Table IV and Table V, 45.75% of tweets which have unsounded code strings at the end of them are replies while 35.54% of all the tweets are replies. 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 5 shows the cumulative relative frequency distribution of

- the length of all the tweets (excluding retweets),
- the length of tweets (excluding retweets) which 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 6, 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 5 and Figure 6 show unsounded code strings at the end of the tweets are mainly 3–4 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 messages are mainly 3–4 multibyte characters long when they are used for smooth communications with particular persons.

Next, we discuss unsounded code strings at the end of normal tweets and replies, individually. Figure 7 shows the cumulative relative frequency distribution of the length of all the normal tweets, the length of normal tweets which 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 8 shows the cumulative relative frequency distribution of the length of all the replies, the length of replies which have unsounded code

 ISBN: 978-1-61208-285-1

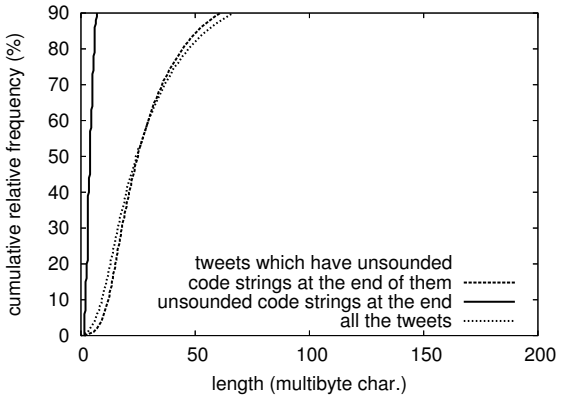

Fig. 5. The cumulative relative frequency distribution of the length of (1) all the tweets, (2) tweets which have unsounded code strings at the end of them, and (3) unsounded code strings at the end of them.

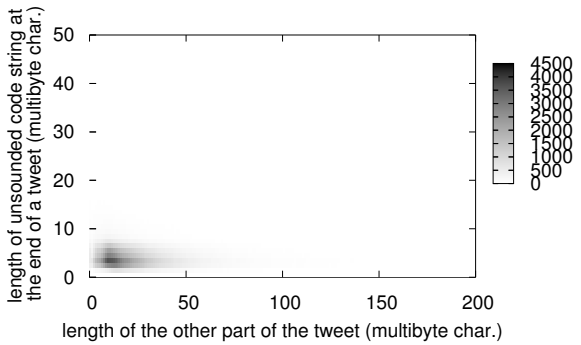

Fig. 6. The heatmap which shows the association between the length of unsounded code string at the end of answers and the other part of the tweets.

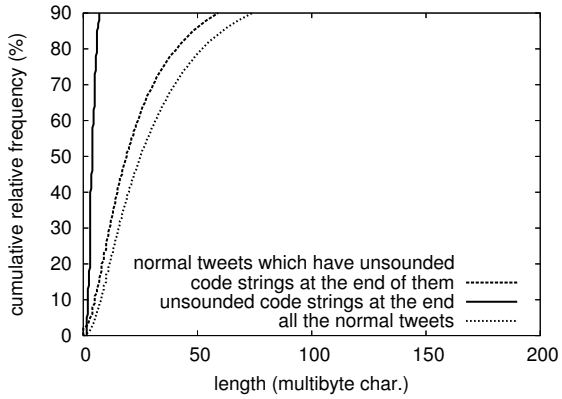

Fig. 7. The cumulative relative frequency distribution of the length of (1) all the normal tweets, (2) normal tweets which have unsounded code strings at the end of them (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

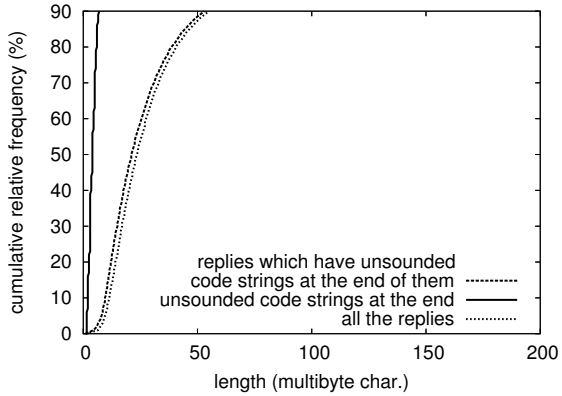

Fig. 8. The cumulative relative frequency distribution of the length of (1) all the replies, (2) replies which have unsounded code strings at the end of them (excluding unsounded code strings at the end of them), and (3) unsounded code strings at the end of them.

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 8, there are few short replies, especially less than 5 multibyte long. It is because each reply includes "@username". Also, as shown in Figure 8, the length of replies which 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 7, the length of normal tweets which 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 there are many normal tweets each of which was sent to general public, not to a particular person. We think a message to general public, not to a particular person, tend to be long because we intend to avoid unnecessary misunderstanding. On the other hand, a message to a particular person is sometimes short. As a result, the distribution of the length of all the normal tweets shifts to longer ranges than the length of normal tweets which have unsounded code strings at the end of them.

V. 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

 ISBN: 978-1-61208-285-1

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 3–4 multibyte characters long. Furthermore, 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.

In this study, we analyzed and compared unsounded code strings only in answers in Yahoo! chiebukuro and tweets in twitter. However, it is not enough to obtain general knowledge about unsounded code strings. It is because both of Yahoo! chiebukuro and twitter have character length limits: Yahoo! chiebukuro has a minimum character length limit, on the other hand, twitter has a maximum character length limit. As a result, we intend to analyze unsounded code strings in a computer aided communication media which has no character length limit.

REFERENCES

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

[2] S. Fahlman. (2012) Smiley:30 years old and never looked happier! [Online]. Available: http://www.cs.cmu.edu/smiley/

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] *Distribution of "Yahoo! Chiebukuro" data*, National Institute of Informatics, 2007. [Online]. Available: http://www.nii.ac.jp/cscenter/idr/yahoo/tdc/chiebukuro_e.html [retrieved: May, 2013]

 ISBN: 978-1-61208-285-1