Analysis of Emoticons in a Japanese Twitter Corpus

Hiroshi Yamaguchi
The University of Aizu
Tsuruga, Ikki-machi,
Aizuwakamatsu, Fukushima, Japan
+81-242-37-2664
m5201115@u-aizu.ac.jp

Maxim Mozgovoy
The University of Aizu
Tsuruga, Ikki-machi,
Aizuwakamatsu, Fukushima, Japan
+81-242-37-2664
mozgovoy@u-aizu.ac.jp

ABSTRACT

This paper describes relation of emoticon and emotion in the text. We used a Japanese Twitter corpus that was annotated with eight emotions based on Ekman and Plutchik emotion categories. We analyze the relation between emoticons and emotions, and briefly discuss obtained results. We are convinced that this work will be useful for researcher interested in emotion analysis of the texts.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis.

General Terms

Algorithms.

Keywords

Twitter, emotions.

1. INTRODUCTION

Emoticons express emotion of a writer texts. We use emoticons in various online contexts, such as e-mailing, chats, and social networking services. An emoticon consists of a string of characters and punctuation symbols such as "(　;　;　)". There is a large number of emoticons used in the Internet for communication, and some emoticons may correspond to several emotions.

In this research we use a Japanese Twitter corpus annotated with six emotions, suggested by Ekman [2] and Plutchik [3]: "anger", "disgust", "sadness", "surprise", "fear" and "happiness", and two two extra emotions, "pride" and "embarrassment" [1]. We manually created the list of all emoticons found in this corpus to establish the link between emotions and emoticons.

Our corpus consists of 4338 tweets. 2215 of them (51.0%) are marked as having emotional value, and 791 of them (18.2%) are marked as containing emoticons. Emotions in

the corpus are distributed very unevenly. For example, the most frequent emotion is "happiness" comprising nearly 40% in this corpus, whereas the least frequent emotion is "fear" comprising nearly 2%.

2. EMOTICON ANALYSIS

To establish the link between emotions and emoticons, we manually created the list of all emoticons found in our corpus. We counted how many times each emotion is marked with a particular emotion. For example if a certain emoticon is marked with the emotion of happiness and pride, it gets one "happiness score" and "pride score". When the same emoticon appears many times, its score is increased. If the same emoticon appears having another emotion, it will also receive the score of that emotion category.

Next, we calculated emotion percentage, showing the range of emotions associated with a particular emoticon, as follows:

$$Emotion \ percentage \ = \ Single \ emotion \ score \ / \ Total \ emotion \ score$$

A score of a single emotion is one of the eight emotion scores such as "happiness score" or "pride score". Total emotion score is the total score of each emoticon. In other words, the sum of "happiness score", "pride score" and so on. The emotion percentage describes the weighted range of emotions that each emoticon may have. It is useful when we discuss the links between individual emotions and emoticons. To understand the contexts of emoticon use in texts, we performed the following process. First, we calculated normalized emotion percentages:

$$Normalized \ percentage \ of \ emotion \ = \ Largest \ percentage \ among \ emotions \ / \ Percentage \ of \ emotion$$

For each emoticon we then calculate a ratio between the highest and the second highest normalized percentage of emotion. The resulting value represents emotional confidence showing the degree of ambiguity of each emoticon. If the ratio is high, it means the emoticon is used almost always in the same emotional context. If the ratio is low, it means that the emoticon is widely used in different contexts, so the emoticon tends to be ambiguous.

The Table 1 shows emotional confidence values of the most frequent emoticons in the corpus. The most frequent emotion is almost always "happiness", but the second frequent

emotion change greatly across emoticons.

The Table 2 describes ten emoticons with the lowest emotional confidence values. Again, the most probable emotion is "happiness".

Table 1: Emotional confidence values for most frequent emoticons in the corpus

Emoticon	Fre	Largest	Second	Ratio
\(^o^)/	40	happiness	surprise	1.9
^^	39	happines	pride	2.9
(^o^)	30	happiness	pride	6.0
( )	28	happiness	pride	6.2
(^^)	23	happiness	pride	5.8
(^O^)/	21	happiness	sadness, pride	8.3
( )	20	sadness	surprise	4.4
(^-^)	17	happiness	surprise	7.7
( )	16	happiness	pride	6.0
(^ ^)	15	happiness	surprise	2.1

Table 2: Emoticons with the lowest emotional confidence

Emoticon	Largest	Second	Ratio
( )	happiness	pride	1.2
><	sadness	happiness	1.2
(-_-)	sadness	happiness	1.3
( )/	happiness	sadness	1.3
f^_^;)	happiness	surprise	1.3
(;)	surprise	sadness	1.5
(^o^	happiness	pride	1.5
(((o(* *)o)))	pride	happiness	1.5
^^;	sadness happiness pride embarrassment	disgust	1.5
((((;( ))))))	sadness	surprise	1.5

3. DISCUSSION

The results of corpus analysis are summarized in the Table 1 and Table 2. If a certain emoticon has a low emotional confidence value, it does not correspond to a specific emotion. Table 2 provides examples of such emoticons that can be used in a variety of emotional contexts.

"Happiness" is the most used emotion in the corpus. "Happiness" is a broad emotional category that is related to "joy", "satisfaction", "love" and other positive emotions, which can explain high frequency of "happiness".

In the Table 2, the most ambiguous emoticon "(
) " represents both "happiness" and "pride". Both emo-

Table 3: Emoticons with the highest emotional confidence

Emoticon	Largest	Second	Ratio
(; ;)	sadness	happiness	12.5
(^^)/	happines	pride	11.0
(^ ^)	happiness	anger	9.0
(^O^)/	happiness	sadness, pride	8.3
(^-^)	happiness	surprise	7.6
( )	happiness	anger, pride	7.5
(* *)	happiness	surprise	7.5
(#^-^#)	happiness	pride	7.0
(*^^*)	happiness	surprise	6.5
( )	happiness	pride	6.2

tions are positive, so this result does not necessarily mean ambiguity.

However, the second ambiguous emoticon "><" represents both "sadness" "happiness". These are completely opposite emotions, so the given emoticon can be used in different contexts.

Unfortunately, most of ambiguous emoticons occur rarely in the corpus, so it is difficult to judge whether they are truly ambiguous or their emotional scores are unreliable. One interesting result is "^^;". This emoticon has high emotional confidence values of "sadness", "happiness", "pride" and "embarrassment". This emoticon can be widely used in both positive and negative contexts.

4. CONCLUSIONS

In this research, we discussed the relation between emoticons and emotions in the text. Some emoticons are used for different emotional meanings in texts. Sometimes an emoticon may represent opposite emotions, depending on the context. Emoticons do not necessarily represent emotions originally associated with them. Furthermore, emoticons complement the emotions not reflected in the words.

The results of this research are quite unbalanced due to biased emotional categorization. Especially, there are nearly 40% of "happiness"-marked emoticons in our corpus. In the future, we will need re-mark the corpus to get a more balanced emotional classification and thus obtain more reliable results.

5. REFERENCES

[1] A. Danielewicz-Betz, H. Kaneda, M. Mozgovoy, and M. Purgina, " Creating English and Japanese Twitter corpora for emotion analysis" , International Journal of Knowledge Engineering, 2015, vol. 1(2), pp. 120-124.

[2] P. Ekman, " Basic Emotions," in Handbook of Cognition and Emotion, T. Dalgleish and M. Power, Eds., John Wiley & Sons, 1999, pp. 45-60.

[3] R. Plutchik, " The nature of emotions," American Scientist, vol. 89, no. 4, 2001, pp. 344-350.