Development of Emoticon Database for Affect Analysis in Japanese

Michal Ptaszynski*, Pawel Dybala*, Radoslaw Komuda†, Rafal Rzepka* and Kenji Araki*

*Department of Information Science and Technology
Hokkaido University, Sapporo 060-0814 Japan
Tel: +81-11-706-7389, Fax: +81-11-709-6277
E-mail: {ptasyznski,paweldybala,komuda,kabura,araki}@media.eng.hokudai.ac.jp

†Faculty of Education Sciences
Nicolaus Copernicus University, ul. Fredry 6/8, Torun 87-100 Poland
Tel: +48-56-611-3110, Fax: +48-56-611-3125
komuda@stud.umk.pl

***Abstract*— In this paper we present our work on creating a database of emoticons – face marks widely used to convey emotions in text-based online communication. The database is created by gathering emoticons from numerous dictionaries of face marks and online jargon. The inconsistencies in emotion classification provided by various dictionaries are solved by processing them with an affect analysis system developed previously. Having the emoticon database annotated automatically this way, we extract patterns from it patterns of semantic areas of emoticons, such as "eyes" and "mouths". Finally, we perform annotation of the semantic areas based on co-occurrence statistics and the theory of kinesics.**

I. INTRODUCTION

This paper presents a research aiming to create a system analyzing emoticons in fully automatic manner. One of the main steps in steps to fulfill that was to create a coherent database of emoticons. To do that we first gather emoticons from Internet web sites that contain emoticon dictionaries. However, these dictionaries, although robust and rich in emotion expression, are usually grouped according to subjective naming criteria. We unify the naming using a previously developed affect analysis system and gather only the groups with naming coherent with scientific classification of emotions applied in our research. Further, the regrouped emoticons are divided into semantic areas, such as representations of mouths and eyes. Finally the semantic areas are annotated based on co-occurrence statistics and the theory of kinesics.

II. DEFINITION OF EMOTICON

In this research we define emoticon as a one-line string of symbols containing at least one set of semantic areas (mouth [M] and eyes [E_L], [E_R]), emoticon borders [B_1], [B_2], and additional areas [S_1] - [S_4]. Although we allow part of the set to be of empty value. See Table 1 for details.

III. THEORY OF KINESICS

The word *kinesics*, as defined by Vargas [4], refers to all non-verbal behavior related to movement, such as postures, gestures and facial expressions and today is the anthropological term for body language. It is studied as an important part of nonverbal (or "iconic") communication along with paralanguage (e.g. voice modulation) and proxemics (e.g. social distance). The term was first used by Birdwhistell in 1952 [5], who founded the theory of kinesics in the fifties and developed it further till seventies [6]. The theory assumes that non-verbal behavior is used in everyday communication systematically and can be studied similarly to language. A minimal part distinguished in kinesics is a *kineme* - the smallest set of body moves containing a certain meaning, e.g. raising eyebrows, or moving eyes upward in face movements. Birdwhistell developed a set of kinemes which he used in annotation of body movements in his research.

A. Emoticons in the View of Kinesics

One of the current applications of kinesics is in annotation of affect display in psychology to determine which emotion is represented by which body movement or face expression. Emoticons can be considered as representations of body language in online text-based communication. This is also confirmed by similarities between kinemes and the semantic areas of emoticons defined in this research. This suggests that the reasoning applied in kinesics can be applied as well to the analysis of emoticons.

Using this reasoning we based our analysis of emotive information conveyed in emoticons on annotations of the particular semantic areas grouped in an automatically constructed emoticon database.

IV. DATABASE OF EMOTICONS

B. Resource Collection

To create a coherent database of emoticons and its semantic areas we first needed a collection of raw emoticons. These were extracted from seven online emoticon dictionaries available on seven popular Web pages dedicated to emoticons: Face-mark Party, Kaomojiya, Kaomoji-toshokan, Kaomoji-café, Kaomoji Paradise, Kaomojisyo and Kaomoji Station[1]. The all of those dictionaries are easily accessible from the Internet and provide a large collection of popular and emoticons.

[1] Respectively: http://www.facemark.jp/facemark.htm, http://kaomojiya.com/, http://www.kaomoji.com/kao/text/, http://kaomoji-cafe.jp/, http://rsmz.net/kaopara/, http://matsucon.net/material/dic/, http://kaosute.net/jisyo/kanjou.shtml

Table 1. Examples of emoticons with different numbers of sets of semantic areas; [M] – mouth; [E_L], [E_R] – eyes; [B_1], [B_2] - emoticon borders; [S_1] - [S_4] - additional areas;

No. of sets	Emoticon	S_1	B_1	S_2	E_L	M	E_R	S_3	B_2	S_4	...
1	ヾ(。·ω·)ノ	ヾ	(	.	·	ω	·	N/A	)	ノ	
1	(——;)		(	N/A	—	N/A	—	;	)		
					SET 01						SET 02
2	(^^)人(^^)		(	N/A	^	N/A	^	N/A	)	人	(^^)
2	☆- (●≧ ▿)人(▿ ≦ ●)- ☆	☆-	(	●	≧	▿	N/A	N/A	)	人	(▿ ≦ ●)- ☆
					SET 01		SET 02		SET 03		SET 04
4	(∇ ˙○)皿´★)ω´☆)∀ ´●)				(∇ ˙○)		皿´★)		ω´☆)		∀ ´●)

C. *Unification of Database Naming*

The data in every dictionary is divided into numerous categories, such as "greetings", "affirmations", "actions", "hobby", or "expressing emotion". The number of categories however and their nomenclature is far from unification. Every dictionary provides its own category number and naming. To solve this problem we processed all of the category names with an affect analysis system, in which one of the procedures is to classify the words according to what emotion type they express [7]. The system uses a coherent classification of emotions based on Nakamura's emotive expressions dictionary developed after a long time study on words describing emotional states in Japanese [8]. The names of categories from online emotion collections which revealed some emotional characterization were re-grouped according to Nakamura's classification of emotions. Finally the emoticons from those collections were extracted from the Web pages. This way we extracted a large number of 11,416 emoticons. However, since some emoticons could appear in more than one collection form the seven, we performed a filtering to extract only the unique ones. The number of emoticons after the filtering was 10,157 (89%). This means that most of emoticons appearing in all seven collections were unique.

D. *Extraction of Semantic Areas*

After collecting the sufficient database of emoticons divided into emotion classes according to the coherent emotion classification, we performed an extraction of all semantic areas appearing in the unique emoticons. The extraction was done according to the definition of an emoticon presented above. Firstly, we defined the possible emoticon borders and extracted all unique triplets of semantic areas for combined eyes and mouth together (E_LME_R). From those triplets we extracted mouths (M) and pairs of eyes (E_L,E_R). Finally, having extracted the triplets E_LME_R and defined the emoticon borders we extracted all existing additional areas (S_1,…,S_4). All unique areas of this kind were summarized in order according to their co-occurrence in the database.

V. Database Statistics

The number of unique combined areas of E_LME_R triplets was 6,142. The number of unique areas representing pairs of eyes (E_L,E_R) was 1,920. The number of unique mouth areas (M) was 1,654.

VI. Conclusions and Future Works

In this paper we presented a short description of a database of emoticons to be used in further research on development of a fully automatic emoticon analysis system. The database contains over ten thousand of unique emoticons collected from the Internet. These emoticons are automatically distributed into emotion classes with the use of previously developed affect analysis system. Finally, the emoticons are divided into semantic areas, like mouths or eyes and the hit-rate statistics of all their co-occurrence was calculated. The division of emoticons into semantic areas is based on Birdwhistell's [5,6] idea of kinemes as minimal meaningful elements in body language, which he described in his theory of kinesics.

The database in its shape as described in this paper will be used in the emoticon analysis system. We will determine the possible emotion affiliations for every element in the database using Spearman's correlation test. As the database contains over ten thousand of emoticons and several thousands of elements for each unique semantic area, in our assumption, the system based on this database will be capable to automatically annotate potential emotion types to any emoticon. There is finite number of semantic areas used by users in emoticons generated during online communication. We believe the database described here is sufficient enough to cover most of the possibilities. If this proves to be true, with the use of semantic area databases the system will potentially be capable of emotion classification of original emoticons generated by the users creatively and not appearing in the database as unique emoticons.

After implementation, we plan to evaluate the system's performance in such areas like: 1) emoticon detection in a sentence (using *kappa* value); 2) emoticon extraction from a sentence (using balanced F-score); 3) division of emoticon into semantic areas (using accuracy); 4) emotion classification of emoticons (using balanced F-score for each of ten emotion types distinguished in this research).

Acknowledgment

This research is partially supported by a Research Grant from the Nissan Science Foundation and The GCOE Program founded by Japan's Ministry of Education, Culture, Sports, Science and Technology. The first author would like to thank Jacek Maciejewski for his invaluable help in putting the idea of this system into reality.

References

[1] N. Suzuki and K. Tsuda, "Express Emoticons Choice Method for Smooth Communication of e-Business", *KES 2006*, Part II, LNAI 4252, pp. 296 - 302, 2006.

[2] D. Derks, A.E.R. Bos, J. von Grumbkow, "Emoticons and social interaction on the Internet: the importance of social context", *Computers in Human Behavior*, 23, pp. 842-849, 2007.

[3] K.C. Chiu, "Explorations in the Effect of Emoticon on Negotiation Process from the Aspect of Communication", Master's Thesis, Department Information Management, National Sun Yat-sen University, 2007.

[4] M. F. Vargas, *Louder than Words: An Introduction to Nonverbal Communication*. Ames: Iowa State UP, 1986.

[5] R. L. Birdwhistell, *Introduction to kinesics: an annotation system for analysis of body motion and gesture*. University of Kentucky Press, 1952.

[6] Birdwhistell, R. 1970. Kinesics and Context. University of Pennsylvania Press, Philadelphia.

[7] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki, "Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum", In *Proceedings of The Conference of the Pacific Association for Computational Linguistics 2009 (PACLING-09)*, pp. 223-228, 2009.

[8] Nakamura, A. *Kanjo hyogen jiten* [Dictionary of Emotive Expressions] (in Japanese). Tokyodo Publishing, Tokyo. 1993.