Research on Emoticons: Review of the Field and Proposal of Research Framework

Michal Ptaszynski †　Rafal Rzepka ‡　Kenji Araki ‡　Yoshio Momouchi §

† JSPS Research Fellow / High-Tech Research Center, Hokkai-Gakuen University
ptaszynski@hgu.jp

‡ Graduate School of Information Science and Technology, Hokkaido University
{kabura,araki}@media.eng.hokudai.ac.jp

§ Department of Electronics and Information Engineering,
Faculty of Engineering, Hokkai-Gakuen University
momouchi@eli.hokkai-s-u.ac.jp

Abstract

Emoticons are string of symbols representing body language in text-based communication. In Natural Language Processing (NLP) emoticons have been considered as unnatural language entities. We argue that, in over 40-year-long history of text based communication, emoticons have gained a status of an indispensable means of support for text based messages. This makes them fully a part of Natural rather than Unnatural Language Processing (UNLP). We argue further that the reason the emoticons have been considered as a part of UNLP lies in the lack of sufficient methods for the analysis of emoticons. We propose including emoticon processing in a set of frequent language processing challenges. We mention our state of the art system for extraction and analysis of Japanese emoticons.

1 Introduction

The term "Unnatural Language Processing" (UNLP), as roughly defined for the needs of Baidu UNLP Contest[1] in 2010, refers to a subfield of NLP dealing with language phenomena which cannot be captured by conventional language processing methods[2]. UNLP defined this way[3] includes such problems as informal expressions, typos, emoticons, onomatopoeia or unknown words. This paper focuses on emoticons. We claim that emoticons are far from being unnatural entities in language and mention some empirical proofs for this claim. We notice further that the reason for the emoticons to have been included in UNLP lies in the lack of sufficient methodology for emoticon analysis. We propose making emoticon processing in a frequent NLP challenge and set a number of problems to be solved within it. As a start point we mention several systems including our state of the art system for extraction and analysis of Japanese emoticons.

2 Definition of Emoticons

Emoticons are representations of body language in text-based messages, where the communication channel is limited to transmission of letters and punctuation marks.

It is not certain when the first emoticon in the history was used, however, different sources point to many interesting discoveries. The oldest known reference[4] is to Abraham Lincoln's speech from 1862, where he used a mark looking like a smiley face ";)". Although there is some doubt on whether it is a deliberately used emoticon, or a typo, the mark is used in a humorous context (after a short annotation "applause and laughter"), which supports the emoticon thesis. The first known typographical emoticons annotated with emotion classes, such as "joy", or "melancholy", appeared probably in the U.S. satirical magazine *Puck*[5] in 1881 (see Figure 1).

In the digital era some of the first widely used emoticons were the ones emerged on PLATO, a system for assisted university coursework [3]. As for the emoticons known today, it is assumed that the first ones were introduced in 1982 by Scott Fahlman of Carnegie Mellon University on a Computer Science BBS[6], from where they spread to Usenet and later to the Internet.

Emoticons have been used in online communication for many years and their numbers have developed depending on the language of use, letter input system, or the kind of community they are used in. They can be roughly divided into three types: A) Western one-line type; B) Eastern one-line type; and C) Multi-line ASCII art type. Western emoticons are known for being rotated by 90 degrees, such as ":-)" (smiling face), or ":-D" (laughing

[1] http://www.baidu.jp/unlp/

[2] Definition after Hagiwara on "UnNatural Language Processing Blog", http://blog.lilyx.net/

[3] The term is also defined differently much earlier as an insufficiency for explaining natural language phenomena by computer-based logic or programming languages in general (for details see [1] and [2]).

[4] "Is That an Emoticon in 1862?". *The New York Times*. 2009-01-19. http://cityroom.blogs.nytimes.com/2009/01/19/hfo-emoticon/

[5] *Puck*, No. 212, p. 65, 30 March 1881.

[6] http://www.cs.cmu.edu/ sef/Orig-Smiley.htm

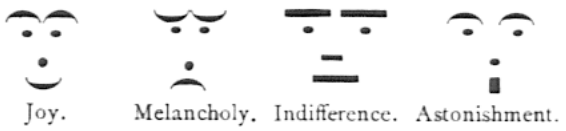

Figure 1: Emoticons presented in the *Puck* magazine.

face). They are usually made of two to four characters and are of a relatively small number. Multi-line ASCII art type emoticons, on the other hand, consist of a number of characters written in several, or even up to several dozens of lines. When looked at from a distance, they make up a picture, often representing a face or a posture. Finally, Eastern emoticons, in contrast to the Western ones are usually unrotated and represent faces or gestures from a point of view easily comprehensible to the reader. Some examples are: " (^○^) " (laughing face) or " (^_^) " (smiling face). They arose in Japan, where they are called *kaomoji*, in the 1980s and since then have been developed in a number of online communities. They are made up of three to over twenty characters written in one line and consist of a representation of at least one face or posture, up to a number of different face-marks.

3 Research on Emoticons - Review

Research on emoticons has developed in two general streams. Firstly, social sciences and communication studies have investigated the effects of emoticons on social interaction. Secondly, in NLP much effort has been put into generating and analyzing emoticons in order to improve computer-related text-based communication and contribute to fields like Computer-Mediated Communication or Human-Computer Interaction.

3.1 Emoticons in Social Sciences

As for social sciences, there are several examples worth mentioning. Ip [4] investigated the impact of emoticons on affect interpretation in Instant Messaging. She concluded that the use of emoticons helps the interlocutors in conveying their emotions during online conversation. Wolf [5] showed further, in her study on newsgroups, that there are significant differences in the use of emoticons by men and women. Derks et al. [6] investigated the influence of social context on the use of emoticons in Internet communication. Finally, linguistic analysis of student chat conversations done by Maness [7] proved emoticons as an important means of online communication.

A thorough research showing the importance of emoticons in communication was presented by Ptaszynski in 2006 [17]. He performed a study on emotive expressions used online and included emoticons as one of such expressions. Firstly, he performed a linguistic analysis of a part of a robust online forum *2channel*[7] to find out which types of emotive expressions appear most often. Secondly, Ptaszynski performed a survey on 110 people (48 women, 62 men of different age groups). In the survey he asked about the types of expressions the participants use to express their feelings when communicating online. Both, the survey and the linguistic analysis showed that emoticons are the second, after direct lexical expressions, most often used type of expressions of emotions (see Figure 2). Moreover, on the lists of ten most popular particular expressions, emoticons appeared most often, 4 times for positive emotions and 5 times for negative emotions (for details see Ptaszynski, 2006 [17], page 90).

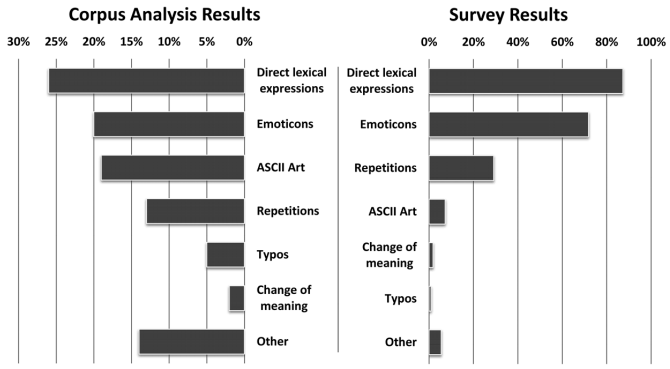

Figure 2: Results presented by Ptaszynski [17], page 92. Left: Percentage of each type of emotive expression used on *2channel*; Right: Popularity of each emotive expression type among survey participants (graphs simplified).

All of the above research is important and prove that emoticons appear frequently in online conversation, and often are necessary for the communication to take place.

3.2 Emoticons in NLP

Two practical applications of research on emoticons in the field of Natural Language Processing are to generate and analyze emoticons. One of the first significant attempts to emoticon generation, was done by Nakamura et al. [8]. They used a Neural Networks-based algorithm to learn a set of emoticon areas (mouths, faces, etc.) and used them later in a dialog agent. Unfortunately, the lack of a firm formalization of the semantic areas made the choice of emoticons eventually random, and the final performance far from ideal. This was one of the reasons for abandoning the idea of emoticon areas as base elements for emoticon-related systems. From that time most of the research on emoticon generation focused on preprogrammed emoticons [9, 10, 11]. In our research [18] we revived the idea of exploiting the emoticon areas, however, not in the research on emoticon generation, but in emoticon extraction and analysis.

There have been several attempts to analyze emoticons or use them to detect user emotions in sentences. For ex-

[7] www.2ch.net/

ample, Reed [12] showed that the use of emoticons can be useful in sentiment classification. Yang et al. [13] made an attempt to use emoticons as seeds to automatically build a lexicon of emotional expressions. Both research focus on preprogrammed Western-type emoticons.

As for attempts to analyze Eastern-type emoticons, there have been three significant ones.

Kernel Method for Emoticon Extraction The system for extraction and classification of emoticons with kernel methods was proposed by Tanaka et al. [14]. They used popular tools for processing sentences in Japanese, a POS tagger ChaSen[8] and a Support Vector Machine-based chunker, *yamcha*[9] to chunk sentences and separate parts of speech from "other areas in the sentence", which they defined as potential emoticons. The method was not ideal, as it was unable to deal with input other than a chunkable sentence and in some cases non-emoticon contents could be recognized as a potential emoticon. Even though, in a closed test on a set of prepared sentences, their best result was somewhat high with 85.5% of Precision and 86.7% of Recall. Their method was also significant as it was the first evaluated attempt to extract emoticons from textual input.

N-gram Method for Emoticon Affect Estimation Yamada et al. [15] used statistics of n-grams to determine emotion types conveyed by emoticons. To classify emoticons they used simple statistics of all characters occurring in emoticons without differentiating them into semantic areas. Eventually this caused errors, as some characters were calculated as "eyes", although they represented "mouths", etc. However, the accuracy of their method was still somewhat high, from 76% to 83%.

CAO is a system for analysis of emoticons in Japanese online communication, developed in our research [18]. The system extracts emoticons from input and determines the specific emotion types they express. Firstly, it matches the extracted emoticons to a predetermined raw emoticon database containing over ten thousand emoticon samples extracted from the Web and annotated automatically. The emoticons, for which emotion types could not be determined using only this database, are automatically divided into semantic areas representing "mouths" or "eyes". These areas are automatically annotated according to their co-occurrence in the database. The evaluation, performed on both training and test sets, confirmed the system's capability to sufficiently detect and extract emoticons, analyze their semantic structure, and estimate the potential emotion types they express. The system achieved nearly ideal scores (over 95%), outperforming the previous emoticon analysis systems.

[8] http://chasen-legacy.sourceforge.jp/

[9] http://chasen.org/ taku/software/yamcha/

4 Framework for NLP Research on Emoticons

Hagiwara includes emoticons in Unnatural Language Processing task (see section 1). This puts emoticons in a position of an unnatural entity in language. In our research on emoticons, summarized in previous sections, we got to the contrary conclusions. Emoticons function as generic representations of body language in text-based communication, and are not only natural, but frequent and often necessary entities in natural language used online. This is also proved by a long history of development and use of emoticons, which emerged along with the first computer-mediated communication environments. Moreover, authors were not able to identify any online communication environment NOT using emoticons as a support for text-based messages.

Despite the firm position in communication, the phenomenon of emoticons has not yet had enough attention (e.g., search engines, including Yahoo or Google, are still incapable of detecting and parsing even the simplest emoticons, which influences search results on informal media, like blogs, etc.). To fill this gap, we propose including research on emoticons as a challenge in NLP. To help researchers investigate this topic in the future we present a framework for the research on emoticons consisting of tasks necessary to fulfill within the research on all types of emoticons.

Table 1 presents our proposal of a framework for research on emoticons. It consists of 12 tasks (11 plus one additional) divided into 6 groups. The first two groups indicate emoticon **detection** and **extraction**. Although these two tasks could be performed with the same procedure, for some tasks it is enough only to confirm the presence of an emoticon (e.g., classifying sentences into emotive and non-emotive). Detection is usually simpler and therefore consumes less time and resources. In extraction, the detected emoticon is further stored in memory, which can be used in further analysis. The next task is emoticon **parsing**, or dividing to particular semantic areas. It is useful, when one aspires to deal with human creativity in emoticon generation [8]. We also proved that it can help in emoticon analysis [18]. Task 4 includes **analysis** of the meaning emoticons convey. The most popular task is to analyze affect, although it is also possible to analyze actions the emoticons depict, or other still unchallenged area. However, it has to be remembered that all these represent different dimensions and should be dealt as separate tasks. Emoticon **generation** is one of the most challenging tasks, as it aims to generate original emoticons to match another contents (e.g., sentence, like in [8]). Recently this task has been simplified to preprogrammed emoticons. Final task to perform within this framework is thorough **evaluation** of the system. It can be performed on separate emoticons, or sentences the emoticons appear in.

To create a system capable to thoroughly analyze

Table 1: Framework for NLP Research on Emoticons.

Task to perform with Emoticon	Type of Emoticon (1-line)	
	A (Western)	B (Eastern)
1. Emoticon Detection		
1.1. Input = emoticon?	–	[18]
1.2. In sentence input	[13]	[14, 18]
1.3. In any input	–	[18]
2. Emoticon Extraction		
2.1. From sentence input	–	[14, 18]
2.2. From any input	–	[18]
3. Parsing / Division into semantic areas	–	[8, 18]
4. Semantic Analysis		
4.1. Affect / Sentiment	[12, 13]	[14, 15, 18]
4.2. Actions	[13]	[14]
4.3. Other?	–	–
5. Emoticon Generation	–	[8, 9, 10]
6. Evaluation on		
6.1. emoticons alone	[13]	[14, 15, 18]
6.2. sentences with emoticons	[12]	[8, 9, 10, 18]

emoticons and use them in a way close to human creativity, all of the above tasks need to be included in the research. However, it is difficult to deal with all emoticon types and tasks at once. Therefore in the table we indicated the research that already dealt each task to some extent. We did not however compare the results, as all research used different datasets and evaluation criteria.

In our research, for example, we focused only on emoticons appearing in online communication in Japanese. Therefore we excluded type-A emoticons (see section 2 for description of emoticon types), since they rarely appear in Japanese online communities. We also did not deal with type-C emoticons (ASCII Art), as their multi-line structure makes their analysis to be considered more as a task for image than language processing. This would be the only way for the computer to obtain an impression of the emoticon from a point of view similar to a user looking at the computer screen. However, type-B emoticons we focused on have a large variation of appearance and are still sophisticated enough to express many different meanings.

5 Conclusions

In this paper we presented an interdisciplinary review of research on emoticons, string of symbols representing body language in text-based communication. We showed that although emoticons have been a part of text based communication for a long time, relatively little have been done in NLP to understand this important phenomenon. We suggested to include emoticon processing as frequent NLP challenge and proposed a framework, containing tasks to be included in the future research on emoticons.

Acknowledgements
This research was supported by (JSPS) KAKENHI Grant-in-Aid for JSPS Fellows (22-00358).

References

[1] Oberlander, J., Monaghan, P., Cox, R., Stenning K. and Tobin R., "Unnatural Language Processing: An Empirical Study of Multimodal Proof Styles", *Journal of Logic, Language and Information*, Vol. 8, pp. 363384, 1999.

[2] Alfred V. Aho, "Unnatural Language Processing", Invited Speech on *The Third Workshop on Syntax and Structure in Statistical Translation*, NAACL-HLT 2009, Boulder, Colorado, June 5, 2009.

[3] Brian L. Dear, "Emoticons and smileys emerged on the PLATO system in the 1970s in a unique and different way". 2008-12-17, http://www.platopeople.com/emoticons.html

[4] Amy Ip. The Impact of Emoticons on Affect Interpretation in Instant Messaging, 2002. http://amysmile.com/pastprj/emoticon_paper.pdf

[5] Alecia Wolf. Emotional Expression Online: Gender Differences in Emoticon Use, *CyberPsychology & Behavior*, Vol. 3, No. 5, pp. 827-833, 2004.

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

[7] Jack M. Maness. A Linguistic Analysis of Chat Reference Conversations with 18-24 Year-Old College Students, *The Journal of Academic Librarianship*, Vol. 34, No. 1, pp. 31-38, 2008.

[8] Jinpei Nakamura, Takeshi Ikeda, Nobuo Inui and Yoshiyuki Kotani. Learning Face Mark for Natural Language Dialogue System, *IEEE NLP-KE 2003*, pp. 180-185, 2003.

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

[10] Nobuo Suzuki and Kazuhiko Tsuda. Automatic emoticon generation method for web community, *IADIS International Conference on Web Based Communities 2006 (WBC2006)*, pp. 331-334, 2006.

[11] Kazumasa Takami, Ryo Yamashita, Kenji Tani, Yoshikazu Honma, Shinichiro Goto. Deducing a User's State of Mind from Analysis of the Pictographic Characters and Emoticons used in Mobile Phone Emails for Personal Content Delivery Services, *International Journal On Advances in Telecommunications*, Vol. 2, No, 1, pp. 37-46, 2009.

[12] Jonathon Read. Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification, In *Proc. of the Student Research Workshop ACL-05*, pp. 43-48, 2005.

[13] Changhua Yang, Kevin Hsin-Yih Lin, Hsin-Hsi Chen. Building Emotion Lexicon from Weblog Corpora, In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 133-136, 2007.

[14] Yuki Tanaka, Hiroya Takamura, Manabu Okumura. Extraction and Classification of Facemarks with Kernel Methods, *IUI'05*, January 9-12, 2005, San Diego, California, USA, 2005.

[15] Taichi Yamada, Seiji Tsuchiya, Shingo Kuroiwa and Fuji Ren. Classification of Facemarks Using N-gram, *Int. Conf. on NLP and Knowledge Engineering*, pp. 322-327, 2007.

[16] Masahiro Kawakami. The database of 31 Japanese emoticon with their emotions and emphases, *The human science research bulletin of Osaka Shoin Women's University*,Vol.7,pp.67-82,2008.

[17] Michal Ptaszynski. Boisterous language: Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum '2channel', M.A. Thesis, UAM, Poznan, 2006. Link: http://arakilab.media.eng.hokudai.ac.jp/~ptaszynski/data/2006_11_SHURON_Moeru_gengo.pdf

[18] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka and Kenji Araki, "CAO: A Fully Automatic Emoticon Analysis System Based on Theory of Kinesics", *IEEE Transactions on Affective Computing*, Vol. 1, No. 1, pp. 46-59, 2010.