Annotating Affective Information on 5.5 Billion Word Corpus of Japanese Blogs

Michal Ptaszynski *　Pawel Dybala †　Rafal Rzepka ‡　Kenji Araki ‡　Yoshio Momouchi §

* High-Tech Research Center, Hokkai-Gakuen University
ptaszynski@hgu.jp
† Department of Information and Management Science, Otaru University of Commerce
paweldybala@res.otaru-uc.ac.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

This paper presents our work on annotating a 5.5 billion word corpus of Japanese blogs with affective information. In the annotation we used two systems for affect analysis: ML-Ask for analyzing words and sentences and CAO for analyzing emoticons. The annotated information includes such features as emotive expressions, emotion classes, emoticons, valence and activation. The statistics of those annotations are presented and compared to other available emotion blog corpora.

1 Introduction

There is a lack of large corpora of Japanese language applicable to emotion processing research. Although there are large corpora of newspaper articles, like Mainichi Shinbun Corpus[1], or corpora of classic literature, like Aozora Bunko[2], they are usually unsuitable for research on emotions since spontaneous emotive expressions either appear rarely in these kinds of texts (newspapers), or the vocabulary is not up to date (classic literature). Recently blogs have come into the focus of sentiment and affect analysis [1, 2, 3]. Therefore annotating a large blog corpus with affective information could help filling the lack in corpora applicable to the research on emotions. In this paper we present the first attempt to automatically annotate affect on a large scale corpus.

The outline of the paper is as follows. In section 2 we describe some of the existing emotion corpora. Section 3 contains description of tools we used. In section 4 we evaluate the annotations, present the statistics and compare them to other emotion corpora. Finally, we conclude the paper and mention some future perspectives.

2 Emotion and Blog Corpora

In this section we compare some of the existing emotion corpora. **YACIS** was collected automatically by Maciejewski et al. [4] from the pages of Ameba blog service. It contains 5.6 billion words within 350 million sentences. The compilation process was performed within 3 weeks between 3rd and 24th of December 2009. The corpus contents are stored in XML files preserving the original blog structure (blog post and comments), thanks to which semantic relations between posts and comments are maintained. The size of raw corpus (pure text corpus without any additional tags) is 27.1 gigabytes. YACIS is the corpus we used in this research to annotate different types of affective information. Quan and Ren [1] created a Chinese emotion blog corpus **Ren-CECps1.0**. They collected 500 blog articles from various Chinese blog services, such as sina blog (http://blog.sina.com.cn/) or qq blog (http://blog.qq.com/). The articles were annotated with information like emotion classes, emotive expressions or valence. The syntactic annotations include tokenization and POS tagging. Wiebe et al. [5] created the **MPQA** corpus of news articles. It contains 10,657 sentences in 535 documents (157 annotated). The annotations include emotive expressions, valence, intensity, etc. However, Wiebe et al. focused mostly on sentiment and subjectivity analysis, and they did not include annotations of emotion classes. Hashimoto et al. [2] developed the **KNB** corpus of Japanese blogs. The corpus contains about 67 thousand words in 249 blog articles. Despite its small scale, the corpus proposes a good standard for preparation of blog corpora for sentiment and affect-related studies. It contains all relevant syntactic annotations (POS, dependency parsing, Named Entity Recognition, etc.) and sentiment-related annotations. The corpus was created mostly for the need of sentiment analysis and therefore does not contain any information on spe-

[1] http://www.nichigai.co.jp/sales/ mainichi/mainichi-data.html
[2] http://www.aozora.gr.jp/

Table 1: Comparison of emotion corpora ordered by the amount of annotations.

corpus name	scale (in sentences / docs)	language	annotated affective information: emotion classes (standard)	emotive expressions	emotive/ non-emot.	valence/ activation	emotion intensity	syntactic annotations
YACIS	354 mil. / 13 mil.	Japanese	10 (language and culture based)	○	○	○/○	○	T,POS,L,DP,NER;
Ren-CECps1.0 [1]	12,724 / 500	Chinese	8 (Yahoo! news)	○	○	○/×	○	T,POS;
MPQA [5]	10,657 / 535	English	none (no standard)	○	○	○/×	○	T,POS;
KNB [2]	4,186 / 249	Japanese	none (no standard)	○	×	○/×	×	T,POS,L,DP,NER;
Minato et al. [6]	1,191 / 1	Japanese	8 (chosen subjectively)	○	○	×/×	×	POS;
Aman&Szpak. [3]	5205 / 173	English	6 (face recognition)	○	○	×/×	○	×

cific emotion classes. However, it is annotated with emotion valence for different categories of valence-related expressions in Japanese, such as *emotional attitude* (e.g., "to feel sad about X" [NEG], "to like X" [POS]), *opinion* (e.g., "X is wonderful" [POS]), or *positive/negative event* (e.g., "X broke down" [NEG], "X was awarded" [POS]). Aman and Szpakowicz [3] created a small-scale English blog corpus. They focused not on syntactic, but on affect-related annotations. They were also some of the first to recognize the task of distinguishing between emotive and non-emotive sentences. Finally, Minato et al. [6] collected a 14,195 word / 1,191 sentence corpus. The corpus is a collection of dictionary examples from "A short dictionary of feelings and emotions in English and Japanese" [7].

3 Affect Annotation Tools

Emotive Expression Dictionary [8] is a collection of over two thousand expressions describing emotional states collected manually from a wide range of literature. It was converted into an emotive expression database by Ptaszynski et al. [9]. The dictionary, developed for over 20 years, is a state-of-the art example of a handcrafted emotive expression lexicon. It also proposes a classification of emotions that is said to reflect the Japanese language and culture the most appropriately: 喜 *ki/yorokobi* (joy), 怒 *dō/ikari* (anger), 哀 *ai/aware* (sorrow, sadness, gloom), 怖 *fu/kowagari* (fear), 恥 *chi/haji* (shame, shyness), 好 *kō/suki* (fondness), 厭 *en/iya* (dislike), 昂 *kō/takaburi* (excitement), 安 *an/yasuragi* (relief), and 驚 *kyō/odoroki* (surprise). All expressions in the dictionary are annotated with those emotion classes. The distribution of expressions within emotion classes is represented in table 2.

Table 2: Distribution of separate expressions across emotion classes in Nakamura's dictionary (overall 2100 ex.).

emotion class	nunber of expressions	emotion class	nunber of expressions
dislike	532	fondness	197
excitement	269	fear	147
sadness	232	surprise	129
joy	224	relief	106
anger	199	shame	65

ML-Ask [9] is a keyword-based system for affect annotation in Japanese. It uses a two-step procedure: **1)** specifying whether an utterance is emotive, and **2)** annotating the particular emotion classes in utterances described as emotive. The emotive sentences are detected on the basis of *emotemes*, emotive features like: interjections, mimetic expressions, vulgar language and emotive markers. The examples in Japanese are respectively: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (syntactic morpheme of verb vulgarization) and '!', or '??' (markers indicating emotive engagement). Emotion class annotation is based on Nakamura's dictionary. ML-Ask has been also supported with Contextual Valence Shifters (CVS) [10] (words and phrases like "not", or "never", which change the valence of an evaluative word). The last distinguishable feature of ML-Ask is implementation of Russell's two dimensional affect model [11], in which emotions are represented in two dimensions: valence (positive/negative) and activation (activated/deactivated).

CAO [12] is a system for affect analysis of Japanese emoticons, called *kaomoji*. CAO extracts emoticons from input and determines specific emotions they express. Firstly, it matches the input to a predetermined emoticon database (over ten thousand emoticons). The emoticons, which could not be estimated this way are divided into semantic areas (representations of "mouth" or "eyes"). The areas are automatically annotated according to their co-occurrence in the database. In the annotation process CAO was used as a supporting procedure in ML-Ask to improve the overall performance and add detailed information about emoticons. An example of outputs of ML-Ask and CAO is represented in figure 1.

Figure 1: Output examples for ML-Ask and CAO.

4 Results and Evaluation

Evaluation of Affective Annotations: Firstly, we needed to confirm the performance of affect analysis systems on YACIS. In the evaluation we used a test set created by Ptaszynski et al. [12] for the evaluation of CAO. It consists of thousand sentences randomly extracted from YACIS and manually annotated with emotion classes by 42 layperson annotators in an anonymous survey. There are 418 emotive and 582 non-emotive sentences. We compared the results on those sentences for ML-Ask, CAO (described in detail in [12]), and both systems combined. The results showing accuracy, calculated as a ratio of success to the overall number of samples, are summarized in table 3. The performance of discrimination between emotive and non-emotive sentences of ML-Ask alone was a high 98.8%. As for CAO, it is capable of detecting the presence of emoticons in a sentence, which is partially equivalent to detecting emotive sentences in ML-Ask. The performance of CAO was also high, 97.6%. This was due to the fact that grand majority of emotive sentences contained emoticons. Finally, ML-Ask supported with CAO achieved remarkable 100% accuracy. This was a surprisingly good result, although it must be remembered that the test sample contained only 1000 sentences (less than 0.0003% of the whole corpus). Next we verified emotion class annotations on sentences. The baseline of ML-Ask achieved 73.4% of accuracy. CAO achieved 80.2%. Interestingly, this makes CAO a better affect analysis system than ML-Ask. However, the condition is that a sentence must contain an emoticon. The best result, close to 90%, was achieved by ML-Ask supported with CAO. We also checked the results when only the dimensions of valence and activation were taken into account. ML-Ask achieved 88.6%, CAO nearly 95%. Support of CAO to ML-Ask again resulted in the best score, 97.5%.

Statistics of Affective Annotations: At first we checked the statistics of emotive and non-emotive sentences, and its determinant features (emotemes). There were nearly twice as many emotive sentences than non-emotive (ratio 1.94). This suggests that the corpus is biased in favor of emotive contents, which could be considered as a proof for the assumption that blogs make a good base for emotion related research. When it comes to statistics of each emotive feature (emoteme), the most frequent class were interjections. Second frequent was the exclamative marks class, such as "!", or "??". Third frequent emoteme class was emoticons, followed by endearments. As an interesting remark, emoteme class that was the least frequent was vulgarities. As one possible interpretation of this result we propose the following. Blogs are social space, where people describe their experiences to be read and commented by other people (friends, colleagues). The use of vulgar language could discourage potential readers from further reading, making the blog less popular. Next, we checked the statistics of emotion classes annotated on emotive sentences. The results are represented in table 5. The most frequent emotions were joy (31%), dislike (20%) and fondness (19%), which cover over 70% of all annotations. However, it could happen that the number of expressions included in each emotion class database influenced the number of annotations (database containing more expressions has a higher probability to gather more annotations). Therefore we calculated the correlation between the number of annotations and the number of emotive expressions in each emotion class database using Spearman's rank correlation test. The test revealed no statistically significant correlation between the two types of data, with ρ=0.38.

Comparison with Other Emotion Corpora: Firstly, we compared YACIS with KNB. We compared the ratios of sentences expressing positive to negative valence. The comparison was made for all KNB valence categories separately and as a sum. In our research we do not make additional sub-categorization of valence types, but used in the comparison ratios of sentences with only positive/negative valence and including the sentences which were mostly positive/negative. The comparison is presented in table 6. In KNB for all valence categories except one the ratio of positive to negative sentences was biased in favor of positive sentences. Moreover, for most cases, the ratio was similar to the one in YACIS (around 1.7). Although the numbers of compared sentences differ, the fact that the ratio remains similar across the two different corpora suggests that the Japanese express in blogs more positive than negative emotions. Next, we

Table 3: Evaluation results of ML-Ask and CAO.

	emotive/ non-emotive	emotion classes	2D (valence and activation)
ML-Ask	98.8%	73.4%	88.6%
CAO	97.6%	80.2%	94.6%
ML-Ask+CAO	100.0%	89.9%	97.5%

Table 4: Statistics of emotive sentences.

# of emotive sentences	233,591,502
# of non-emotive sentence	120,408,023
ratio (emotive/non-emotive)	1.94
# of sentences containing emoteme class:	
- interjections	171,734,464
- exclamative marks	89,626,215
- emoticons	49,095,123
- endearments	12,935,510
- vulgarities	1,686,943
ratio (emoteme classes in emotive sentence)	1.39

Table 5: Emotion class annotations with percentage.

emotion class	**# of sentences**	**%**	**emotion class**	**# of sentences**	**%**
joy	16,728,452	31%	excitement	2,833,388	5%
dislike	10,806,765	20%	surprize	2,398,535	5%
fondness	9,861,466	19%	gloom	2,144,492	4%
fear	3,308,288	6%	anger	1,140,865	2%
relief	3,104,774	6%	shame	952,188	2%

Table 6: Comparison of positive and negative sentences between KNB and YACIS.

		positive	negative	ratio
KNB*	emotional attitude	317	208	1.52
	opinion	489	289	1.69
	merit	449	264	1.70
	acceptation or rejection	125	41	3.05
	event	43	63	0.68
	sum	1,423	865	1.65
YACIS**	only	22,381,992	12,837,728	1.74
	only+mostly	23,753,762	13,605,514	1.75

* p<.05, ** p<.01

Table 7: Comparison of number of emotive expressions appearing in three different corpora with the results of Spearman's rank correlation test.

	Minato et al.	YACIS	Nakamura
dislike	355	14,184,697	532
joy	295	22,100,500	224
fondness	205	13,817,116	197
sorrow	205	2,881,166	232
anger	160	1,564,059	199
fear	145	4,496,250	147
surprise	25	3,108,017	129
	Minato et al. and Nakamura	**Minato et al. and YACIS**	**YACIS and Nakamura**
Spearman's ρ	0.88	0.63	0.25

compared the corpus created by Minato et al. [6]. This corpus was prepared on the basis of an emotive expression dictionary. Therefore we compared its statistics not only to YACIS, but also to the emotive lexicon used in our research [8]. Emotion classes used in Minato et al. differ slightly to those used in our research. For example, they use class name "hate" to describe what in YACIS is called "dislike". Moreover, they have no classes such as "excitement", "relief" or "shame". To make the comparison possible we used only the emotion classes appearing in both cases and unified all class names. The results are summarized in table 7. There was no correlation between YACIS and Nakamura (ρ=0.25), which confirms the results calculated in the previous paragraph. A medium correlation was observed between YACIS and Minato et al. (ρ=0.63). Finally, a strong correlation was observed between Minato et al. and Nakamura (ρ=0.88), which is the most interesting observation. Both Minato et al. and Nakamura are in fact dictionaries of emotive expressions. The fact that they strongly correlate suggests that for the compared emotion classes there could be a general tendency in language to create more expressions to describe some emotions rather than the others (dislike, joy and fondness are often some of the most frequent emotion classes). This phenomenon needs to be verified more thoroughly in the future.

5 Conclusions and Future Work

In this paper we described our attempt to automatically annotate affective information on a large scale blog corpus. The annotations were done on YACIS, over 5.5 billion word corpus of blogs collected from Ameba blog corpus by Maciejewski et al. [4]. The systems used in the annotation process include ML-Ask, a system for affect analysis of utterances and CAO, a system for affect analysis of emoticons. The evaluation on a test sample of annotations showed sufficiently high results. Statistics of the affective annotations were compared to other emotion corpora. The comparison showed similarities in the ratio of expressions of positive to negative emotions on both small and large scale corpora. We also observed a high correlation between two different emotive expression dictionaries. YACIS corpus annotated with affective information can be further applied in research on emotions in language, sentiment and affect analysis. In the near future we also plan to provide a demo viewable online allowing corpus querying for affective information.

Acknowledgments

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

References

[1] Quan, C. and Ren, F. 2010. "A blog emotion corpus for emotional expression analysis in Chinese", *Computer Speech & Language*, Vol. 24, Issue 4, pp. 726-749.

[2] Hashimoto, C., Kurohashi, S., Kawahara, D., Shinzato, K. and Nagata, M. 2011. "Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations" [in Japanese], *Journal of Natural Language Processing*, Vol. 18, No. 2, pp. 175-201.

[3] Aman, S. and Szpakowicz, S. 2007. "Identifying Expressions of Emotion in Text", *LNAI* 4629, pp. 196-205.

[4] Maciejewski, J., Ptaszynski, M., Dybala, P. 2010. "Developing a Large-Scale Corpus for Natural Language Processing and Emotion Processing Research in Japanese", In *Proceedings of the International Workshop on Modern Science and Technology (IWMST)*, pp. 192-195.

[5] Wiebe, J., Wilson, T. and Cardie, C. 2005. "Annotating expressions of opinions and emotions in language", *Language Resources and Evaluation*, Vol. 39, Issue 2-3, pp. 165-210.

[6] Minato, J., Bracewell, D. B., Ren, F. and Kuroiwa, S. 2006. "Statistical Analysis of a Japanese Emotion Corpus for Natural Language Processing", *LNCS* 4114, pp. 924-928.

[7] Hiejima, I. 1995. *A short dictionary of feelings and emotions in English and Japanese*, Tokyodo Publishing.

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

[9] Ptaszynski, M., Dybala, P., Rzepka, R. and Araki, K. 2009. "Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -", In *Proceedings of PACLING-09*, pp. 223-228.

[10] Zaenen, A. and Polanyi, L. 2006. "Contextual Valence Shifters", In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu, J. Wiebe (eds.), Springer Verlag, pp. 1-10.

[11] Russell, J. A. 1980. "A circumplex model of affect", *J. of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161-1178.

[12] Ptaszynski, M., Maciejewski, J., Dybala, P., Rzepka, R., Araki, K. 2010. "CAO: Fully Automatic Emoticon Analysis System", In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 1026-1032.