Automatically Annotating A Five-Billion-Word Corpus of Japanese Blogs for Affect and Sentiment Analysis

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

This paper presents our research on automatic annotation of a five-billion-word corpus of Japanese blogs with information on affect and sentiment. We first perform a study in emotion blog corpora to discover that there has been no large scale emotion corpus available for the Japanese language. We choose the largest blog corpus for the language and annotate it with the use of two systems for affect analysis: ML-Ask for word- and sentence-level affect analysis and CAO for detailed analysis of emoticons. The annotated information includes affective features like sentence subjectivity (emotive/non-emotive) or emotion classes (joy, sadness, etc.), useful in affect analysis. The annotations are also generalized on a 2-dimensional model of affect to obtain information on sentence valence/polarity (positive/negative) useful in sentiment analysis. The annotations are evaluated in several ways. Firstly, on a test set of a thousand sentences extracted randomly and evaluated by over forty respondents. Secondly, the statistics of annotations are compared to other existing emotion blog corpora. Finally, the corpus is applied in several tasks, such as generation of emotion object ontology or retrieval of emotional and moral consequences of actions.

1 Introduction

There is a lack of large corpora for Japanese applicable in sentiment and affect analysis. 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). Although there exist speech corpora, such as Corpus of Spontaneous Japanese[3], which could become suitable for this kind of research, due to the difficulties with compilation of such corpora they are relatively small. In research such as the one by Abbasi and Chen (2007) it was proved that public Internet services, such as forums or blogs, are a good material for affect analysis because of their richness in evaluative and emotive information. One kind of these services are blogs, open diaries in which people encapsulate their own experiences, opinions and feelings to be read and commented by other people. Recently blogs have come into the focus of opinion mining or sentiment and affect analysis (Aman and Szpakowicz, 2007; Quan and Ren, 2010). Therefore creating a large blog-based emotion corpus could help overcome both problems: the lack in quantity of corpora and their applicability in sentiment and affect analysis. There have been only a few small Japanese emotion corpora developed so far (Hashimoto et al., 2011). On the other hand, although there exist large Web-based corpora (Erjavec et al., 2008; Baroni and Ueyama, 2006), access to them is usually allowed only from the Web interface, which makes additional annotations with affective information difficult. In this paper we present the first attempt to automatically annotate affect on YACIS, a large scale corpus of Japanese blogs. To do that we use two systems for affect analysis of Japanese, one for word- and sentence-level affect analysis and another especially for detailed analysis of emoticons, to annotate on the corpus different kinds of affective information (emotive expressions, emotion classes, etc.).

[1] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[2] http://www.aozora.gr.jp/
[3] http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/

The outline of the paper is as follows. Section 2 describes the related research in emotion corpora. Section 3 presents our choice of the corpus for annotation of affect- and sentiment-related information. Section 4 describes tools used in annotation. Section 5 presents detailed data and evaluation of the annotations. Section 6 presents tasks in which the corpus has already been applied. Finally the paper is concluded and future applications are discussed.

2 Emotion Corpora

Research on Affect Analysis has resulted in a number of systems developed within several years (Aman and Szpakowicz, 2007; Ptaszynski et al., 2009c; Matsumoto et al., 2011). Unfortunately, most of such research ends in proposing and evaluating a system. The real world application that would be desirable, such as annotating affective information on linguistic data is limited to processing a usually small test sample in the evaluation. The small number of annotated emotion corpora that exist are mostly of limited scale and are annotated manually. Below we describe and compare some of the most notable emotion corpora. Interestingly, six out of eight emotion corpora described below are created from blogs. The comparison is summarized in Table 1. We also included information on the work described in this paper for better comparison (YACIS).

Quan and Ren (2010) 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/), qq blog (http://blog.qq.com/), etc., and annotated them with a large variety of information, such as emotion class, emotive expressions or polarity level. Although syntactic annotations were simplified to tokenization and POS tagging, this corpus can be considered a state-of-the-art emotion blog corpus. The motivation for Quan and Ren is also similar to ours - dealing with the lack of large corpora for sentiment analysis in Chinese (in our case - Japanese).

Wiebe et al. (2005) report on creating the **MPQA** corpus of news articles. The corpus contains 10,657 sentences in 535 documents[4]. The annotation schema includes a variety of emotion-related information, such as emotive expressions, emotion valence, intensity, etc. However, Wiebe et al. focused on detecting subjective (emotive) sentences, which do not necessarily convey emotions, and classifying them into positive and negative. Thus their annotation schema, although one of the richest, does not include emotion classes.

A corpus of Japanese blogs, called **KNB**, rich in the amount and diversification of annotated information was developed by Hashimoto et al. (2011). It contains 67 thousand words in 249 blog articles. Although it is not a small scale corpus, it developed a certain standard for preparing corpora, especially blog corpora for sentiment and affect-related studies in Japan. The corpus contains all relevant grammatical annotations, including POS tagging, dependency parsing or Named Entity Recognition. It also contains sentiment-related information. Words and phrases expressing emotional attitude were annotated by laypeople as either positive or negative. One disadvantage of the corpus, apart from its small scale, is the way it was created. Eighty one students were employed to write blogs about different topics especially for the need of this research. It could be argued that since the students knew their blogs will be read mostly by their teachers, they selected their words more carefully than they would in private.

Aman and Szpakowicz (2007) constructed a small-scale English blog corpus. They did not include any grammatical information, but focused on affect-related annotations. As an interesting remark, they were some of the first to recognize the task of distinguishing between emotive and non-emotive sentences. This problem is usually one of the most difficult in text-based Affect Analysis and is therefore often omitted in such research. In our research we applied a system proved to deal with this task with high accuracy for Japanese.

Das and Bandyopadhyay (2010) constructed an emotion annotated corpus of blogs in Bengali. The corpus contains 12,149 sentences within 123 blog posts extracted from Bengali web blog archive (http://www.amarblog.com/). It is annotated with face recognition annotation standard (Ekman, 1992).

Matsumoto et al. (2011) created ***Wakamono Kotoba*** (Slang of the Youth) corpus. It contains unrelated sentences extracted manually from Yahoo! blogs (http://blog-search.yahoo.co.jp/). Each sen-

[4]The new MPQA Opinion Corpus version 2.0 contains additional 157 documents, 692 documents in total.

Table 1: Comparison of emotion corpora ordered by the amount of annotations (abbreviations: T=tokenization, POS=part-of-speech tagging, L=lemmatization, DP=dependency parsing, NER=Named Entity Recognition).

corpus name	scale (in sentences / docs)	language	annotated affective information						syntactic annotations
			emotion class standard	emotive expressions	emotive/ non-emot.	valence/ activation	emotion intensity	emotion objects	
YACIS	354 mil. /13 mil.	Japanese	10 (language and culture based)	○	○	○/○	○	○	T,POS,L,DP,NER;
Ren-CECps1.0	12,724/500	Chinese	8 (Yahoo! news)	○	○	○/×	○	○	T,POS;
MPQA	10,657/535	English	none (no standard)	○	○	○/×	○	○	T,POS;
KNB	4,186/249	Japanese	none (no standard)	○	×	○/×	×	○	T,POS,L,DP,NER;
Minato et al.	1,191sent.	Japanese	8 (chosen subjectively)	○	○	×/×	×	×	POS;
Aman&Szpak.	5,205/173	English	6 (face recognition)	○	○	×/×	○	×	×
Das&Bandyo.	12,149/123	Bengali	6 (face recognition)	○	×	×/×	○	×	×
Wakamono Kotoba	4773sentences	Japanese	9 (face recognition + 3 added subjectively)	○	×	×/×	×	×	×
Mishne	?/815,494	English	132 (LiveJournal)	×	×	×/×	×	×	×

tence contains at least one word from a slang lexicon and one word from an emotion lexicon, with additional emotion class tags added per sentence. The emotion class set used for annotation was chosen subjectively, by applying the 6 class face recognition standard and adding 3 classes of their choice.

Mishne (2005) collected a corpus of English blogs from LiveJournal (http://www.livejournal.com/) blogs. The corpus contains 815,494 blog posts, from which many are annotated with emotions (moods) by the blog authors themselves. The LiveJournal service offers an option for its users to annotate their mood while writing the blog. The list of 132 moods include words like "amused", or "angry". The LiveJournal mood annotation standard offers a rich vocabulary to describe the writer's mood. However, this richness has been considered troublesome to generalize the data in a meaningful manner (Quan and Ren, 2010).

Finally, Minato et al. (2006) collected a 14,195 word, 1,191 sentence corpus. The corpus was a collection of sentence examples from a dictionary of emotional expressions (Hiejima, 1995). The dictionary was created for the need of Japanese language learners. Differently to the dictionary applied in our research (Nakamura, 1993), in Hiejima (1995) sentence examples were mostly written by the author of the dictionary himself. The dictionary also does not propose any coherent emotion class list, but rather the emotion concepts are chosen subjectively. Although the corpus by Minato et al. is the smallest of all mentioned above, its statistics is described in detail. Therefore in this paper we use it as one of the Japanese emotion corpora to compare our work to.

All of the above corpora were annotated manually or semi-automatically. In this research we performed the first attempt to annotate a large scale blog corpus (YACIS) with affective information fully automatically. We did this with systems based on positively evaluated affect annotation schema, performance, and standardized emotion class typology.

3 Choice of Blog Corpus

Although Japanese is a well recognized and described world language, there have been only few large corpora for this language. For example, Erjavec et al. (2008) gathered a 400-million-word scale Web corpus **JpWaC**, or Baroni and Ueyama (2006) developed a medium-sized corpus of Japanese blogs **jBlogs** containing 62 million words. However, both research faced several problems, such as character encoding, or web page metadata extraction, such as the page title or author which differ between domains. Apart from the above mentioned medium sized corpora at present the largest Web based blog corpus available for Japanese is **YACIS** or **Y**et **A**nother **C**orpus of **I**nternet **S**entences. We chose this corpus for the annotation of affective information for several reasons. It was collected automatically by Maciejewski et al. (2010) from the pages of Ameba blog service. It contains 5.6 billion words within 350 million sentences. Maciejewski et al. were able to extract only pages containing Japanese posts (pages with legal disclaimers or written in languages other than Japanese were omitted). In the initial phase they provided their crawler, optimized to crawl only Ameba blog service, with 1000 links

```
<doc url="http://ameblo.jp/blog-name/entry-000001.html"
time="2009-12-05 21:11:46" id="2000001">
   <post>
      <s>今日から十月です。</s>
      [Its October from today.]
      <s>なんか、九月はいつもよりアッという間に過ぎたような気がするなぁ。</s>
      [I have a strange feeling September passed faster than usual.]
      ...
   </post>
   <comments>
      <cmt>
         <s>色々と忙しいですね～！</s>
         [Oh, you've been busy, weren't you?]
         ...
      </cmt>
      <cmt>
         <s>お疲れｻﾏです(^O^)</s>
         [Well done! Cheers for good work (^o^)]
         ...
      </cmt>
   </comments>
</doc>
```

Figure 1: The example of YACIS XML structure.

Table 2: General Statistics of YACIS.

# of web pages	12,938,606
# of unique bloggers	60,658
average # of pages/blogger	213.3
# of pages with comments	6,421,577
# of comments	50,560,024
average # of comment/page	7.873
# of words	5,600,597,095
# of all sentences	354,288,529
# of words per sentence (average)	15
# of characters per sentence (average)	77

taken from Google (response to one simple query: 'site:ameblo.jp'). They saved all pages to disk as raw HTML files (each page in a separate file) and afterward extracted all the posts and comments and divided them into sentences. The original structure (blog post and comments) was preserved, thanks to which semantic relations between posts and comments were retained. The blog service from which the corpus was extracted (Ameba) is encoded by default in Unicode, thus there was no problem with character encoding. It also has a clear and stable HTML meta-structure, thanks to which they managed to extract metadata such as blog title and author. The corpus was first presented as an unannotated corpus. Recently Ptaszynski et al. (2012b) annotated it with syntactic information, such as POS, dependency structure or named entity recognition. An example of the original blog structure in XML is represented in Figure 1. Some statistics about the corpus are represented in Table 2.

4 Affective Information Annotation Tools

Emotive Expression Dictionary (Nakamura, 1993) is a collection of over two thousand expressions describing emotional states collected manually from a wide range of literature. It is not a tool *per se*, but

Figure 2: Output examples for ML-Ask and CAO.

Table 3: 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
		sum	**2100**

was converted into an emotive expression database by Ptaszynski et al. (2009c). Since YACIS is a Japanese language corpus, for the affect annotation we needed the most appropriate lexicon for the language. The dictionary, developed for over 20 years by Akira Nakamura, is a state-of-the art example of a hand-crafted emotive expression lexicon. It also proposes a classification of emotions that reflects the Japanese culture: 喜 *ki/yorokobi*[5] (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 one emotion class or more if applicable. The distribution of expressions across all emotion classes is represented in Table 3.

ML-Ask (Ptaszynski et al., 2009a; Ptaszynski et al., 2009c) is a keyword-based language-dependent system for affect annotation on sentences 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, emoticons

[5] Separation by "/" represents two possible readings of the character.

Table 4: 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%

and emotive markers. The examples in Japanese are respectively: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (syntactic morpheme used in verb vulgarization), (^_^) (emoticon expressing joy) and '!', '??' (markers indicating emotive engagement). Emotion class annotation is based on Nakamura's dictionary. ML-Ask is also the only present system for Japanese recognized to implement the idea of Contextual Valence Shifters (CVS) (Zaenen and Polanyi, 2005) (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 (Russell, 1980), in which emotions are represented in two dimensions: valence (positive/negative) and activation (activated/deactivated). An example of negative-activated emotion could be "anger"; a positive-deactivated emotion is, e.g., "relief". The mapping of Nakamura's emotion classes on Russell's two dimensions was proved reliable in several research (Ptaszynski et al., 2009b; Ptaszynski et al., 2009c; Ptaszynski et al., 2010b). With these settings ML-Ask detects emotive sentences with a high accuracy (90%) and annotates affect on utterances with a sufficiently high Precision (85.7%), but low Recall (54.7%). Although low Recall is a disadvantage, we assumed that in a corpus as big as YACIS there should still be plenty of data.

CAO (Ptaszynski et al., 2010b) is a system for affect analysis of Japanese emoticons, called *kaomoji*. Emoticons are sets of symbols used to convey emotions in text-based online communication, such as blogs. CAO extracts emoticons from input and determines specific emotions expressed by them. Firstly, it matches the input to a predetermined raw emoticon database (with over ten thousand emoticons). The emoticons, which could not be estimated with this database are divided into semantic areas (representations of "mouth" or "eyes"). The areas are automatically annotated according to their co-occurrence in the database. The performance of CAO was evaluated as close to ideal (Ptaszynski et al., 2010b) (over 97%). In this research we used CAO as a supporting procedure in ML-Ask to improve the overall performance and add detailed information about emoticons.

Table 5: 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

5 Annotation Results and Evaluation

It is physically impossible to manually evaluate all annotations on the corpus[6]. Therefore we applied three different types of evaluation. First was based on a sample of 1000 sentences randomly extracted from the corpus and annotated by laypeople. In second we compared YACIS annotations to other emotion corpora. The third evaluation was application based and is be described in section 6.

Evaluation of Affective Annotations: Firstly, we needed to confirm the performance of affect analysis systems on YACIS, since the performance is often related to the type of test set used in evaluation. ML-Ask was positively evaluated on separate sentences and on an online forum (Ptaszynski et al., 2009c). However, it was not yet evaluated on blogs. Moreover, the version of ML-Ask supported by CAO has not been evaluated thoroughly as well. In the evaluation we used a test set created by Ptaszynski et al. (2010b) 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 by Ptaszynski et al. (2010b)), and both systems combined. The results showing accuracy, cal-

[6]Having one sec. to evaluate one sentence, one evaluator would need 11.2 years to verify the whole corpus (354 mil.s.).

Table 6: 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%

culated as a ratio of success to the overall number of samples, are summarized in Table 4. The performance of discrimination between emotive and non-emotive sentences of ML-Ask baseline was a high 98.8%, which is much higher than in original evaluation of ML-Ask (around 90%). This could indicate that sentences with which the system was not able to deal with appear much less frequently on Ameblo. 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, since emoticons are one type of features determining sentence as emotive. 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 slightly better results (73.4%) than in its primary evaluation (Ptaszynski et al., 2009c) (67% of balanced F-score with P=85.7% and R=54.7%). 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: 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, which includes punctuation marks suggesting emotive engagement (such as "!", or "??"). Third frequent emoteme class was emoticons, followed by endearments. As an interesting remark, emoteme class that was the least frequent were 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 6. The most frequent emotions were joy (31%), dislike (20%) and fondness (19%), which covered 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 many expressions has higher probability to gather more annotations). Therefore we verified if there was a correlation between the number of annotations and the number of emotive expressions in each emotion class database. The verification was based on Spearman's rank correlation test between the two sets of numbers. The test revealed no statistically significant correlation between the two types of data, with ρ=0.38.

Table 7: 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

Comparison with Other Emotion Corpora: Firstly, we compared YACIS with KNB. The KNB corpus was annotated mostly for the need of sentiment analysis and therefore does not contain any

Table 8: Comparison of number of emotive expressions in three different corpora including ratio within this set of emotions and results of Spearman's rank correlation test.

	Minato et al.	YACIS	Nakamura
dislike	355 (26%)	14,184,697 (23%)	532 (32%)
joy	295 (21%)	22,100,500 (36%)	224 (13%)
fondness	205 (15%)	13,817,116 (22%)	197 (12%)
sorrow	205 (15%)	2,881,166 (5%)	232 (14%)
anger	160 (12%)	1,564,059 (3%)	199 (12%)
fear	145 (10%)	4,496,250 (7%)	147 (9%)
surprise	25 (2%)	3,108,017 (5%)	129 (8%)
	Minato et al. and Nakamura	**Minato et al. and YACIS**	**YACIS and Nakamura**
Spearman's ρ	0.88	0.63	0.25

information on specific emotion classes. However, it is annotated with emotion valence for different categories valence is expressed 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]). 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 subcategorization of valence types, but used in the comparison ratios of sentences in which the expressed emotions were of only positive/negative valence and including the sentences which were mostly (in majority) positive/negative. The comparison is presented in table 7. 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, including the ratio taken from the sums of sentences, the ratio was similar to the one in YACIS (around 1.7). Although the numbers of compared sentences differ greatly, 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 compared the corpus created by Minato et al. (2006). 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 (see section 4 for details). Emotion classes used in Minato et al. differ slightly to those used in our research (YACIS and Nakamura's dictionary). 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 8. There was no correlation between YACIS and Nakamura (ρ=0.25), which confirms the results calculated in 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. However, the dictionaries were collected in different times (difference of about 20 years), by people with different background (lexicographer vs. language teacher), based on different data (literature vs. conversation) assumptions and goals (creating a lexicon vs. Japanese language teaching). The only similarity is in the methodology. In both cases the dictionary authors collected expressions considered to be emotion-related. The fact that they correlate so strongly suggests that for the compared emotion classes there could be a 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.

6 Applications

6.1 Extraction of Evaluation Datasets

In evaluation of sentiment and affect analysis systems it is very important to provide a statistically reliable random sample of sentences or documents as a test set (to be further annotated by laypeople). The larger is the source, the more statistically reliable is the test set. Since YACIS contains 354 mil. sentences in 13 mil. documents, it can be considered sufficiently reliable for the task of test set extraction, as probability of extracting twice the same sentence is close to zero. Ptaszynski et al. (2010b) already used YACIS to randomly extract a 1000 sentence sample and used it in their evaluation of emoticon analysis system. The sample was also used in this research and is described in more detail in section 5.

6.2 Generation of Emotion Object Ontology

One of the applications of large corpora is to extract from them smaller sub-corpora for specified tasks. Ptaszynski et al. (2012a) applied YACIS for their task of generating an robust emotion object ontology. They used cross-reference of annotations of emotional information described in this paper and syntactic annotations done by Ptaszynski et al. (2012b) to extract only sentences in which expression of emotion was proceeded by its cause, like in the example below.

彼女に振られたから悲しい...
Kanojo ni furareta ***kara*** *kanashii...*
Girlfriend DAT dump PAS CAUS sad ...
I'm sad **because** my girlfriend dumped me...

The example can be analyzed in the following way. Emotive expression (*kanashii*, "sad") is related with the sentence contents (*Kanojo ni furareta*, "my girlfriend dumped me") with a causality morpheme (***kara***, "**because**"). In such situation the sentence contents represent the object of emotion. This can be generalized to the following meta-structure,

$$O_E \quad CAUS \quad X_E,$$

where O_E=[Emotion object], $CAUS$=[**causal form**], and X_E=[expression of emotion].

The cause phrases were cleaned of irrelevant words like stop words to leave only the object phrases. The evaluation showed they were able to extract nearly 20 mil. object phrases, from which 80% was extracted correctly with a reliable significance. Thanks to rich annotations on YACIS corpus the ontology included such features as emotion class (joy, anger, etc.), dimensions (valence/activation), POS or semantic categories (hypernyms, etc.).

6.3 Retrieval of Moral Consequence of Actions

Third application of the YACIS corpus annotated with affect- and sentiment-related information has been in a novel research on retrieval of moral consequences of actions, first proposed by Rzepka and Araki (2005) and recently developed by Komuda et al. (2010)[7]. The moral consequence retrieval agent was based on the idea of Wisdom of Crowd. In particular Komuda et al. (2010) used a Web-mining technique to gather consequences of actions applying causality relations, like in the research described in section 6.2, but with a reversed algorithm and lexicon containing not only emotional but also ethical notions. They cross-referenced emotional and ethical information about a certain phrase (such as "To kill a person.") to obtain statistical probability for emotional ("feeling sad", "being in joy", etc.) and ethical consequences ("being punished", "being praised", etc.). Initially, the moral agent was based on the whole Internet contents. However, multiple queries to search engine APIs made by the agent caused constant blocking of IP address an in effect hindered the development of the agent.

The agent was tested on over 100 ethically-significant real world problems, such as "killing a man", "stealing money", "bribing someone", "helping people" or "saving environment". In result 86% of recognitions were correct. Some examples of the results are presented in the Appendix on the end of this paper.

7 Conclusions

We performed automatic annotation of a five-billion-word corpus of Japanese blogs with information on affect and sentiment. A survey in emotion blog corpora showed there has been no large scale emotion corpus available for the Japanese language. We chose YACIS, a large-scale blog corpus and annotated it using two systems for affect analysis for word- and sentence-level affect analysis and for analysis of emoticons. The annotated information included affective features like sentence subjectivity (emotive/non-emotive) or emotion classes (joy, sadness, etc.), useful in affect analysis and information on sentence valence/polarity (positive/negative) useful in sentiment analysis obtained as generalizations of those features on a 2-dimensional model of affect. We evaluated the annotations in several ways. Firstly, on a test set of thousand sentences extracted and evaluated by over forty respondents. Secondly, we compared the statistics of annotations to other existing emotion corpora. Finally, we showed several tasks the corpus has already been applied in, such as generation of emotion object ontology or retrieval of emotional and moral consequences of actions.

[7] See also a mention in *Scientific American*, by Anderson and Anderson (2010).

Acknowledgments

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

References

Ahmed Abbasi and Hsinchun Chen. "Affect Intensity Analysis of Dark Web Forums", Intelligence and Security Informatics 2007, pp. 282-288, 2007

Saima Aman and Stan Szpakowicz. 2007. "Identifying Expressions of Emotion in Text". In *Proceedings of the 10th International Conference on Text, Speech, and Dialogue (TSD-2007)*, Lecture Notes in Computer Science (LNCS), Springer-Verlag.

Michael Anderson and Susan Leigh Anderson. 2010. "Robot be Good", *Scientific American*, October, pp. 72-77.

Dipankar Das, Sivaji Bandyopadhyay, "Labeling Emotion in Bengali Blog Corpus ? A Fine Grained Tagging at Sentence Level", Proceedings of the 8th Workshop on Asian Language Resources, pages 47?55, 2010.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta. 2008. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", Kluwer Academic Publishers, Netherlands.

Marco Baroni and Motoko Ueyama. 2006. "Building General- and Special-Purpose Corpora by Web Crawling", In *Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application*, www.tokuteicorpus.jp/result/pdf/2006_004.pdf

Jürgen Broschart. 1997. "Why Tongan does it differently: Categorial Distinctions in a Language without Nouns and Verbs." *Linguistic Typology*, Vol. 1, No. 2, pp. 123-165.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale and Mark Johnson. 2000. "BLLIP 1987-89 WSJ Corpus Release 1", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43

Paul Ekman. 1992. "An Argument for Basic Emotions". *Cognition and Emotion*, Vol. 6, pp. 169-200.

Irena Srdanovic Erjavec, Tomaz Erjavec and Adam Kilgarriff. 2008. "A web corpus and word sketches for Japanese", *Information and Media Technologies*, Vol. 3, No. 3, pp.529-551.

Katarzyna Głowińska and Adam Przepiórkowski. 2010. "The Design of Syntactic Annotation Levels in the National Corpus of Polish", In *Proceedings of LREC 2010*.

Peter Halacsy, Andras Kornai, Laszlo Nemeth, Andras Rung, Istvan Szakadat and Vikto Tron. 2004. "Creating open language resources for Hungarian". In *Proceedings of the LREC*, Lisbon, Portugal.

Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato and Masaaki Nagata. 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.

Ichiro Hiejima. 1995. *A short dictionary of feelings and emotions in English and Japanese*, Tokyodo Shuppan.

Paul J. Hopper and Sandra A. Thompson. 1985. "The Iconicity of the Universal Categories 'Noun' and 'Verbs'". In *Typological Studies in Language: Iconicity and Syntax*. John Haiman (ed.), Vol. 6, pp. 151-183, Amsterdam: John Benjamins Publishing Company.

Daisuke Kawahara and Sadao Kurohashi. 2006. "A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis", *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 176-183.

Radoslaw Komuda, Michal Ptaszynski, Yoshio Momouchi, Rafal Rzepka, and Kenji Araki. 2010. "Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions", *Int. Journal of Computational Linguistics Research*, Vol. 1 , Issue 3, pp. 155-163.

Taku Kudo and Hideto Kazawa. 2009. "Japanese Web N-gram Version 1", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T08

Vinci Liu and James R. Curran. 2006. "Web Text Corpus for Natural Language Processing", In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 233-240.

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.

Kazuyuki Matsumoto, Yusuke Konishi, Hidemichi Sayama, Fuji Ren. 2011. "Analysis of Wakamono Kotoba Emotion Corpus and Its Application in Emotion Estimation", *International Journal of Advanced Intelligence*, Vol.3,No.1,pp.1-24.

Junko Minato, David B. Bracewell, Fuji Ren and Shingo Kuroiwa. 2006. "Statistical Analysis of a Japanese Emotion Corpus for Natural Language Processing", *LNCS* 4114.

Gilad Mishne. 2005. "Experiments with Mood Classification in Blog Posts". In *The 1st Workshop on Stylistic Analysis of Text for Information Access*, at *SIGIR 2005*, August 2005.

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

Jan Pomikálek, Pavel Rychlý and Adam Kilgarriff. 2009. "Scaling to Billion-plus Word Corpora", In *Advances in Computational Linguistics*, *Research in Computing Science*, Vol. 41, pp. 3-14.

Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki. 2009. "A System for Affect Analysis of Utterances in Japanese Supported with Web Mining", *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol. 21, No. 2, pp. 30-49 (194-213).

Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki. 2009. "Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States". In *Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp. 1469-1474.

Michal Ptaszynski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2009. "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 (PACLING-09)*, pp. 223-228.

Michal Ptaszynski, Rafal Rzepka and Kenji Araki. 2010a. "On the Need for Context Processing in Affective Computing", In *Proceedings of Fuzzy System Symposium (FSS2010)*, Organized Session on Emotions, September 13-15.

Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2010b. "CAO: Fully Automatic Emoticon Analysis System", In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 1026-1032.

Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2012a. "A Robust Ontology of Emotion Objects", In *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 719-722.

Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2012b. "Annotating Syntactic Information on 5.5 Billion Word Corpus of Japanese Blogs", In *Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 385-388.

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

Rafal Rzepka, Kenji Araki. 2005. "What Statistics Could Do for Ethics? - The Idea of Common Sense Processing Based Safety Valve", AAAI Fall Symposium on Machine Ethics, *Technical Report FS-05-06*, pp. 85-87.

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

Peter D. Turney and Michael L. Littman. 2002. "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus", National Research Council, Institute for Information Technology, *Technical Report ERB-1094*. (NRC #44929).

Masao Utiyama and Hitoshi Isahara. 2003. "Reliable Measures for Aligning Japanese-English News Articles and Sentences". *ACL-2003*, pp. 72-79.

Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. "Annotating expressions of opinions and emotions in language". *Language Resources and Evaluation*, Vol. 39, Issue 2-3, pp. 165-210.

Theresa Wilson and Janyce Wiebe. 2005. "Annotating Attributions and Private States", In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II*, pp. 53-60.

Annie Zaenen and Livia Polanyi. 2006. "Contextual Valence Shifters". In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu, J. Wiebe (eds.), Springer Verlag, Dordrecht, The Netherlands, pp. 1-10.

Appendix. Examples of emotional and ethical consequence retrieval.

emotional conseq.	results	score	ethical conseq.	results	score
SUCCESS CASES					
"To hurt somebody."					
anger	13.01/54.1	0.24	penalty/	4.01/7.1	0.565
fear	12.01/54.1	0.22	punishment		
sadness	11.01/54.1	0.2			
"To kill one's own mother."					
sadness	9.01/35.1	0.26	penalty/	5.01/5.1	0.982
surprise	6.01/35.1	0.17	punishment		
anger	5.01/35.1	0.14			
"To steal an apple."					
surprise	2.01/6.1	0.33	reprimand/	3.01/3.1	0.971
anger	2.01/6.1	0.33	scold		
"To steal money."					
anger	3.01/9.1	0.33	penalty/punish.	3.01/6.1	0.493
sadness	2.01/9.1	0.22	reprimand/sco.	2.01/6.1	0.330
"To kill an animal."					
dislike	7.01/23.1	0.3	penalty/	36.01/45.1	0.798
sadness	5.01/23.1	0.22	punishment		
"To drive after drinking."					
fear	6.01/19.1	0.31	penalty/punish.	24.01/36.1	0.665
"To cause a war."					
dislike	7.01/15.1	0.46	illegal	2.01/3.1	0.648
fear	3.01/15.1	0.2			
"To stop a war."					
joy	6.01/13.1	0.46	forgiven	1.01/1.1	0.918
surprise	2.01/13.1	0.15			
"To prostitute oneself."					
anger	6.01/19.1	0.31	illegal	12.01/19.1	0.629
sadness	5.01/19.1	0.26			
"To have an affair."					
sadness	10,01/35.1	0.29	penalty/punish.	8.01/11.1	0.722
anger	9.01/35.1	0.26			
INCONSISTENCY BETWEEN EMOTIONS AND ETHICS					
"To kill a president."					
joy	2.01/4.1	0.49	penalty/	2.01/2.1	0.957
likeness	1.01/4.1	0.25	punishment		
"To kill a criminal."					
joy	8.01/39.1	0.2	penalty/	556/561	0.991
excite	8.01/39.1	0.2	punishment		
anger	7.01/39.1	0.18			
CONTEXT DEPENDENT					
"To act violently."					
anger	4.01/11.1	0.36	penalty/punish.	1.01/2.1	0.481
fear	2.01/11.1	0.18	agreement	1.01/2.1	0.481
NO ETHICAL CONSEQUENCES					
"Sky is blue."					
joy	51.01/110,1	0.46	none	0	0
sadness	21.01/110,1	0.19			