原著論文

A System for Affect Analysis of Utterances in Japanese Supported with Web Mining

Michal PTASZYNSKI *・Pawel DYBALA *・Wenhan SHI *・Rafal RZEPKA *・Kenji ARAKI *

We propose a method for affect analysis of textual input in Japanese supported with Web mining. The method is based on a pragmatic reasoning that emotional states of a speaker are conveyed by emotional expressions used in emotive utterances. It means that if an emotive expression is used in a sentence in a context described as emotive, the emotion conveyed in the text is revealed by the used emotive expression. The system ML－Ask（Emotive Elements / Expressions Analysis System）is constructed on the basis of this idea. An evaluation of the system is performed in which two evaluation methods are compared. To choose the most objective evaluation method we compare the most popular method in the field and a method proposed by us. The proposed evaluation method was shown to be more objective and revealed the strong and weak points of the system in detail. In the evaluation experiment ML-Ask reached human level in recognizing the general emotiveness of an utterance（0.83 balanced F－score）and 63% of human level in recognizing the specific types of emotions. We support the system with a Web mining technique to improve the performance of emotional state types extraction. In the Web mining technique emotive associations are extracted from the Web using co－occurrences of emotive expressions with morphemes of causality. The Web mining technique improved the performance of the emotional states types extraction to 85% of human performance.

Keywords：Affect analysis, Emotiveness, Analysis of emotiveness, Web mining, Evaluation methods.

1．Introduction

Scientists have been fascinated by emotions for centuries. There are remarkable works trying to describe emotions, such as the ones by Darwin（Darwin, 1872）, or many others（Izard, 1977；Frijda, 1987；Ortony, Clore and Collins, 1988）. However, for a long time emotions were treated rather as an idol to worship, worthy of the attention of thinkers but not material or tangible enough to be accurately described in detail or processed by machines. Recent years have brought research on emotions into focus in Computer Science and Artificial Intelligence, and its sub-fields like Natural Language Processing（Picard, 1997；Nakayama, Eguchi and Kando, 2004）. The subjectivity of emotions however, drives researchers into a corner of ambiguity and often becomes an impediment to research in this field. However, we believe automatic analysis of emotions, narrowed down to specified bounds, should give results comparable to those of humans.

Technological development has led to the creation of a new dimension of communication, man-machine communication（MMC）(Geiser, 1990), where a machine is one part. A rush in development of talking robots and conversational systems（e.g. Higuchi et al., 2008）indicates that the functional implementation in our lives of agents like intelligent car navigation systems（Takahashi et al., 2003）or talking furniture（Hase et al., 2007）has already become a current process, and a need for humanized interfaces in MMC grows rapidly.

One of the most important cognitive human behaviors present in everyday communication is expressing and understanding emotions. People make many decisions under the influence of their own emotions or the emotional states of others（Schwarz, 2000；Young 2006）. Therefore one of the current issues in fields like agent development or communication and computer science is to produce methods for effectively recognizing and analyzing users' emotional states.

* Hokkaido University, Graduate School of Information Science and Technology

Recent introductions in our lives of technologies such as 3G, ubiquitous networks or wireless LAN made the Internet, a social phenomenon only a few years ago (Morris and Ogan, 1996), an indispensable "everyday article". There was a rapid increase of fields using Internet resources as an object of research, like information or opinion retrieval (Singhal, 2001). The Web is becoming a determinant of human commonsense (Rzepka et al., 2006). Within the last few years there were several trials to retrieve information concerning human emotions and attitudes (Abbasi and Chen, 2007) from the Internet. However, there have not been many attempts to use Web mining for affect analysis.

Furthermore, although being a rather young discipline of study, the research on mechanical processing of emotions, including recognition and analysis of affect, has been gathering popularity of researchers since being initiated only a little over ten years ago (Picard, 1995). A number of scientists have proposed their ideas on how to recognize emotions automatically, presenting higher or lower results. However, on what grounds the results have been achieved has been a problem preferably left unsaid till now. Even after over ten years of development of the field, there have been no significant or reliable ideas on how to objectively evaluate emotion recognition methods. Unreliable measures might put the fairness of the results in question or doubt. However, it is understandable that many scientists keep using questionable methods for evaluation since methods said to have a potential to clearly verify the agreement of emotional states with its realizations in language, like functional Magnetic Resonance Imaging (fMRI), are usually laborious time consuming and expensive. Aside form difficulties in performing the fMRI experiments a difficult problem is also how much the performing of such an experiment in itself would influence the participants' emotional states. For example, a subject might feel nervous because of being plugged into the fMRI apparatus, which would obviously influence the results of a questionnaire about the subject's emotional state. Therefore there was a need to work out a method of evaluation not having a distinguishable influence on the results, and at the same time being easy to perform and more objective than the methods normally used in the field today. Therefore, in this paper, after presenting our method for affect recognition from textual input, we also propose a fair and thorough method of evaluation for systems like ours, analyzing the emotive content of utterances.

The outline of the paper is as follows. First we describe our approach and introduce notions used commonly in this paper. Secondly, we propose definitions and classification of emotions used in this research. Next, a baseline system, constructed on the introduced notions, is proposed, followed by a Web mining technique to support it. After that, we describe some of the main problems considering evaluation of affect analysis systems and our proposal for an evaluation method potentially capable of solving these problems. We evaluate our system using one of the most popular evaluation methods in the field today and the method proposed by us. The results are summarized and finally concluding remarks are described and future work discussed.

2. Affect Analysis

Affect analysis is a sub-field of information extraction. Its goal is to estimate human emotional states in communication and humanizing machines by implementing methods for recognition of users' emotional states. Popular methods for analyzing affect include analyzing emotions from facial expressions or voice (Kang et al., 2000). However, since emotions are strongly context dependent, most of the semantic content of expressing emotions is ignored in such research.

Therefore, we decided to analyze the affect of textual representations of utterances. Furthermore, our method of affect analysis uses the advantages of Internet resources widely available today.

2.1 Our Approach to Affect Analysis

There are different dimensions in which emotions can be analyzed, such as psychological, visual, vocal, linguistic, social, neurobiological, etc. Unfortunately, although it would be desirable to combine these fields for creating a multidimensional emotion analyzer, the development of research in each of these fields is still not sophisticated enough. Therefore, to create a firm ground for similar research in the future, we decided to limit our approach to emotions to the textual surface form of speech analysis, since it is language that makes human emotions so context-dependent and thus difficult to process only by their visual or phonetic manifestations.

Many attempts to analyze the emotive aspects of text end in a failure. Read (Read, 2004) helplessly strived against the lack of an appropriate emotive lexicon and the unsystemized classification of emotion types, in the end achieving an accuracy of 33%[1]. Alm et al. (Alm, Roth, and Sproat, 2005) achieved an accuracy of 69%. They too wrestled head on with the lack of emotive databases, but also with an inappropriate evaluation corpus consisting of children's stories full of ambiguities arising from mixing styles and means of expression (e.g. dialogues mixed with narrative style and, descriptions). Finally, Wu et al. (Wu, Chuang and Lin 2006) achieved 72% general average accuracy, but ran into a problem with ambiguity of emotional rules. Another problem is the evaluation of such systems. In much research the means of evaluation are unstandardized, performed with a small number of evaluators (Matsumoto et al., 2007 ; Tokuhisa et al., 2008) with crucial parts of the evaluation performed by the researchers themselves (Wu, Chuang and Lin, 2006).

Therefore we made the following assumptions to specify what we mean by affect analysis. The research will be conducted on the basis of textual utterances. Compared to using the large and quickly rising number of textual corpora, it is time consuming and uneconomical to gather a relevant corpus for the analysis of vocal and visual dimensions. Moreover, neglecting semantic meaning in such research leads to ambiguities in emotional state determination since some realizations of non-verbal communication, like high pitch or arousal can convey different emotional states. Restricting ourselves to a textual corpus thus gives a green light to concentrate on emotiveness realized in the semantic surface of the text. The analysis will be conducted on dialogue-like utterances. This limits appearance of utterances realizing descriptive or poetic function of language[2], such as literature and poetry. This approach relates to the future goal, which is to implement the system into a conversational system like Higuchi's (Higuchi et al., 2008) to make the utterance understanding and generation more natural. The language we will work with is Japanese. The analysis of affect is performed in three steps : 1. Determining the emotiveness of an utterance, or finding whether an utterance is emotive or not ; 2. Finding how strong the emotions conveyed in the utterance are, or setting the emotive value ; 3. If the sentence is emotive, determining what types of emotions are conveyed in the utterance.

3. Linguistic Approach to Emotions

In an everyday conversation there are different linguistic and paralinguistic means used to inform other interlocutors of emotional states. The elements of speech used to convey emotive meaning in sentences make up the feature of language called emotiveness (Stevenson, 1963 ; Kamei, Kouno and Chino, 1996) and the function of language describing them is called the emotive function of language (Jakobson, 1960 ; Bühler, 1978).

The emotive function of language is realized lexically in Japanese through such parts of speech as the following : exclamations (Beijer, 2002 ; Ono, 2002), hypocoristics (endearments) (Potts and Kawahara, 2004), vulgar language (Crystal, 1989) or mimetic expressions (*gitaigo*) (Baba, 2001, 2003). A key role in expressing emotions is also played by the lexicon of words describing states of emotions (Nakamura, 1993). On the borderline between verbality and nonverbality we can talk about elements of language such as intonation, voice modulation or tone of voice. In the written text these are usually represented symbolically by exclamation marks, multiple usage of question marks, and so on. Nonverbal elements realizing emotive language are body language, with all its components, like gestures, facial expressions, eye contact, or pose (Darwin, 1872 ; Hall, 1966 ; Argyle, 1967). However, in conversation systems the communication channel is limited to transmission of signals encoded in lines of letters, punctuation marks, symbols, etc. Therefore, for analysis of emotiveness in conversation systems we need to agree to a compromise of restrictions in the communication channel and focus the analysis on its linguistic manifestations, like exclamation marks or ellipsis.

[1] All results rounded off to the nearest whole percentage unit.

[2] For full description of functions of language see Jakobson 1960.

4. Definition and Classification of Emotions

In the field of affect analysis there is a tendency to forge definitions and classifications of emotions for the needs of a particular research project. This might arouse doubts as to whether the classification was not done to match the results. To avoid this problem and to contribute to standardization of criteria, we decided to use definitions and classifications of emotions based on the most thorough research in this matter currently available for the Japanese language (Nakamura, 1993) rather than create our own.

4.1 Definition of Emotions

We use the general definition of emotions proposed by Nakamura (1993), who defines them as every temporary state of mind or emotional state evoked by experiencing different sensations. This definition is complemented by Beijer's (2002) definition of emotive utterances, which he describes as "every utterance in which the speaker in question is emotionally involved, and in which this involvement is linguistically expressed by means of intonation or by the use of performative expressions."

4.2 Classification of Emotions

Nakamura (1993), after a thorough study on emotions in Japanese, proposed a classification of emotions into 10 types - said to be the most appropriate for the Japanese language. That is : 喜 (*ki, yorokobi*– "joy, delight"), 怒 (*do, ikari*–"anger"), 哀 (*ai, aware* – "sorrow, sadness"), 怖 (*fu, kowagari*–"fear"), 恥 (*chi, haji*–"shame, shyness, bashfulness"), 好 (*kou, suki*– "liking, fondness"), 厭 (*en, iya*–"dislike, detestation"), 昂 (*kou, takaburi*–"excitement"), 安 (*an, yasuragi*– "relief") and 驚 (*kyou, odoroki*–"surprise, amazement").

5. Emotive Elements / Emotive Expressions Analysis System (ML-Ask)

Based on the linguistic approach towards emotions described above as well the definition and classification of emotions, we constructed a system working in three steps :

1) Analyzing the general emotiveness of an utterance;
2) Calculating the emotive value representing the strength of the conveyed emotions;
3) Recognizing the particular emotion types.

The analysis of emotiveness in the proposed system is based on Ptaszynski's idea of two part classification of realizations of emotions in language (Ptaszynski 2006). The classification says that in language there are:

1. *Emotive elements* - indicating that emotions have been conveyed, but not expressing specific emotions ; or, more precisely, expressing different emotions depending on the context of the sentence. For example : すげぇ *sugee* (great!), ワクワク *wakuwaku* (heart pounding), –やがる –*yagaru* (fu**ing do sth -vulgarization of a verb), –ちゃん –*chan* (hypocoristic name suffix)
2. *Emotive expressions* - not always used in emotive utterances, but in emotive utterances describing emotional states. For example : 愛情 *aijou* (love), 悲しむ *kanashimu* (feel sad), むかつく *mukatsuku* (get angry), むしずが走る *mushizu ga hashiru* (give one the creeps), 歓天喜地 *kantenkichi* (delight larger than Heaven and Earth), 嬉しい *ureshii* (happy), 怖い *kowai* (scary);

However for some of the emotive elements the ambiguity of emotional affiliation is restricted to only some emotions and there are emotive elements that are also used frequently as expressions of emotions of a certain type. Based on this classification, Ptaszynski et al. proposed a method for finding emotive elements in text (Ptaszynski, Dybala, Shi, Rzepka and Araki, 2007 ; Ptaszynski, Dybala, Rzepka and Araki, 2008a). In a textual input utterance provided by a user the three features (emotiveness, emotive value and emotional state) are determined by cross-referencing top-down determined databases of emotive elements and emotive expressions in speech.

We constructed these databases for Japanese, based on various previous research. The databases of emotive elements are divided into interjections (Nakamura, 1993 ; Oshima-Takane, MacWhinney, Shirai, Miyata and Naka, 1995-1998 ; Tsuchiya, 1999 ; Itoh, Minematsu and Nakagawa, 1999 ; Ono, 2002), emotive mimetics (*gitaigo*) (Nakamura, 1993 ; Oshima-Takane, MacWhinney, Shirai, Miyata and Naka, 1995-1998 ; Baba, 2001, 2003), endearments (Kamei, Kouno and Chino, 1996 ; Potts and Kawahara, 2004), and vulgar vocabulary (Sjöbergh, 2006, 2008), which all belong to the lexical layer of speech, and symbols rep-

Table 1　Examples of sentences with similar semantic meaning, but different emotive meaning

		Non-emotive sentence	Emotive sentence
Example I	Japanese	今日はいい天気です。	**ああ、今日はええ天気だな　！（^o^）**
	Romanized transcription	*Kyō wa ii tenki desu.*	***Aa,*** *kyō wa* ***ee*** *tenki* ***dana ! (^o^)***
	English translation	It is a good weather today.	**Wow, now** today is **a fine weatha'! :D**
Example II	Japanese	彼女は、大きいかさをもってきて、信之介を強く殴った。	**あいつぁ　でっけーかさをもってきやがって、シンちゃんをひでー　ボコボコに　しちまった　！**
	Romanized transcription	*Kanojo wa, ookii kasa wo mottekite, Shinnosuke wo tsuyoku nagutta.*	***Aitsua dekkē*** *kasa wo motteki****yagatte,*** *Shin-****chan*** *wo* ***hidē bokoboko*** *ni* ***shichimatta !***
	English translation	She brought a large umbrella and strongly hit Shinnnosuke.	**That slut lugged a huge** umbrella with her and **beat the crap out of Shin**.

resenting emotive elements from the non-lexical layer of speech (Kamei, Kouno and Chino, 1996), like exclamation marks, syllable prolongation marks, etc. As a part of the emotive symbols database, we also added an algorithm recognizing emoticons, as these are symbols already widespread and commonly used in everyday Internet communication tools. A few simple examples of sentences without emotive value, and those colored with emotions are given below (see Table 1 for details). The parts of each sentence that constitute its emotiveness are boldfaced.

After analyzing every utterance this way, the system returns a verdict on whether the utterance is emotive and what emotive elements were found in the utterance. In the next step, in the utterances determined as emotive, the system determines the emotive value of the utterance. The emotive value represents the general strength of emotions conveyed in the sentence. The value is given on a scale of 0 to 5. We checked two methods for emotive value calculation - qualitative and quantitative. The former gives 0 for the lack of emotive elements in the utterance and 1 point for each appearance of every type of emotive element (interjection, mimetic expression, endearment, vulgarity or emotive symbol). The latter counts 1 point for every emotive element found in the sentence (but with a maximum value of 5). We compared these two methods and chose the latter as it gave results closer to human evaluators. In the last step, the system determines the specific emotions conveyed in the utterance. This process is done by cross-referencing the extracted emotive elements with the databases of emotive expressions created from Nakamura's collection (Nakamura, 2004) (see Figure 1 for details). This database lists what emotion type each emotive expression belongs to. An example of anlaysis performed by ML-Ask is shown in Table 2.

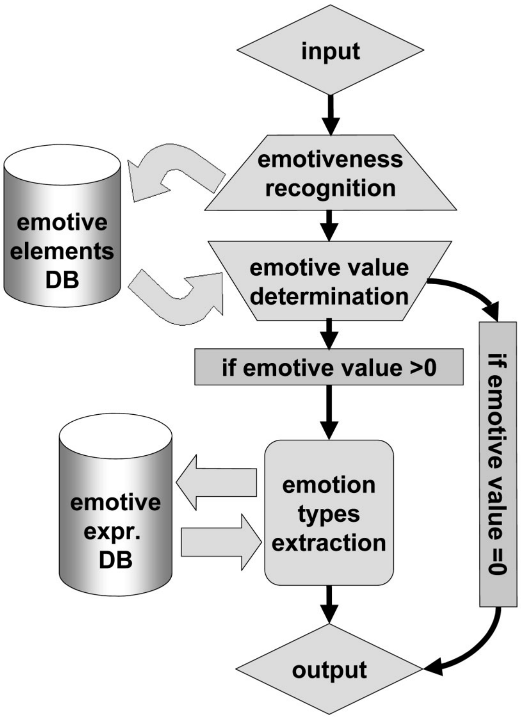


Fig.1　Flow chart of the ML- Ask system

Table 2　An example of analysis performed by ML- Ask

Utterance	*Kono hon saa, sugee kowakatta yo. Maji kowasugi.* (That book, ya know, 'twas a killer. It was just too scary.)
Emotive elements	*saa, sugee, -yo, maji, -sugi* (**Emotive value** = 5)
emotive expressions	*kowai*

Table 3 Example of n‑gram separation from an utterance

Original utterance	*Aa, pasokon ga kowarete shimatta*...(Darn, the PC has broken...)					
longest n-gram (here: hexagram)	*Aa* [interjection]	*pasokon* [noun]	*ga* [particle]	*koware-* [verb]	*te* [verb connector]	*shimau* [perfect form]
pentagram	*pasokon ga koware te shimau*					
tetragram	*Aa, pasokon ga kowareru*					
trigrams	*pasokon ga kowareru*			*koware te shimau*		

6. Web Mining Technique

To improve the extraction of the specific emotion types we apply Shi's Web mining technique for extracting emotive content from the Web based on causality (Shi, 2008). However, since Shi's technique was effective only when a particular phrase to check was extracted manually, we enhanced it with an automatic phrase extraction method based on dividing the sentence into separate n-gram phrases.

6.1 Phrase Extraction From an Utterance

An utterance is first processed by MeCab, a tool for part-of-speech analysis of Japanese (Kudo, 2001). Every element separated by MeCab is treated as a unigram. All the unigrams are grouped into larger n-gram groups preserving their word order in the utterance. The groups are arranged from the longest n-gram (the whole sentence) down to all groups of trigrams. N-grams ending with particles are excluded, since they gave too many ambiguous results in our pre-test phase. An example result of this sentence separating procedure is shown in Table 3.

6.2 Emotion Types Extraction Using the Web

In the web mining the extracted phrase is used as a query to the Google search engine. Moreover, n-grams ending with a verb or an adjective are grammatically adjusted for semantically deeper Web mining. They are modified in line with Yamashita's research on causality morphemes conveying emotive meaning (Yamashita, 1999) independently confirmed experimentally by Shi (2008), as in Table 4. This morpheme modification provides additional query phrases. All the phrases, the extracted ones as well as the modified versions of the ones ending with a verb or an adjective, are queried in Google. A maximum number of 500 snippets for each queried phrase is extracted from the Web and cross-referenced with the database of emotional expressions described in section 5 (see Fig. 3). The emotive expressions

Table 4 Examples of n‑gram modifications for Web mining

	Original n-gram	*pasokon ga koware te shimau*
n-gram phrase adjusting (morpheme modification)	*/ -te /*	*pasokon ga koware te shima- tte*
	/ -to /	*pasokon ga koware te shimau to*
	/ -node /	*pasokon ga koware te shimau node*
	/ -kara /	*pasokon ga koware te shimau kara*
	...	...

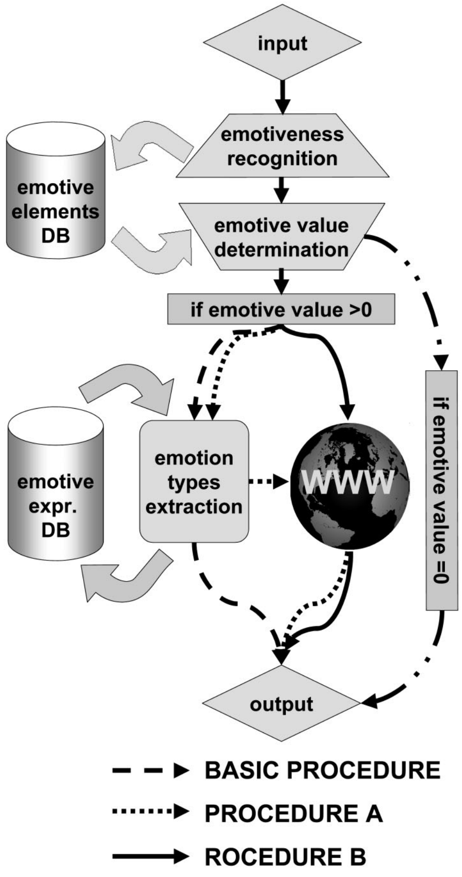


Fig. 2 Flow chart of the ML‑Ask system with 3 procedures for emotion types extraction

Table 5 Example of emotion types extraction from the Web

Sentence: *Aa, pasokon ga kowarete shimatta*...(Darn, the PC has broken…)

Extracted emotion type	Type extracted / all extracted types	Ratio
[fear]	28 / 133	0.210526315789474
[sorrow, sadness]	26 / 133	0.195488721804511
[dislike, detestation]	16 / 133	0.120300751879699
[liking, fondness]	14 / 133	0.105263157894737
[relief]	12 / 133	0.090225563909774
[excitement]	11 / 133	0.082706766917293
[joy, delight]	10 / 133	0.075187969924812
[surprise, amazement]	9 / 133	0.067669172932330
[anger]	5 / 133	0.037593984962406
[shame, shyness, bashfulness]	2 / 133	0.015037593984962

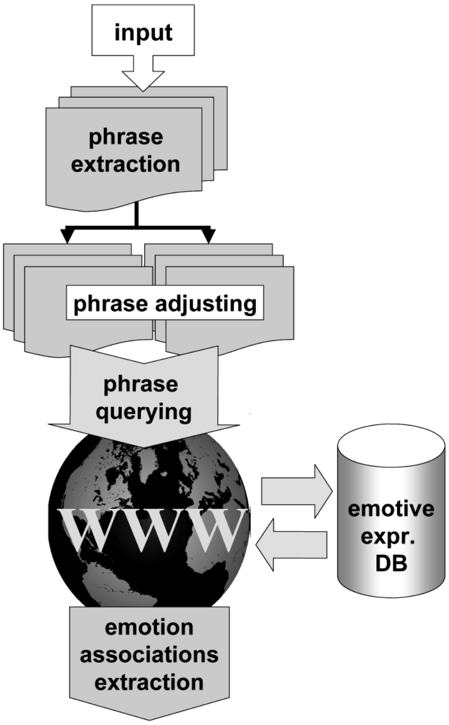


Fig.3 Flow chart of the Web mining technique

extracted from the snippets are added up, and the results for every emotion type are listed in descending order. This way a list of emotions commonsensically associated with the queried sentence is obtained (an example is shown in Table 5).

In the experiment we evaluated the ML-Ask system's performance for the whole three step procedure described in section 5 and compared the results for emotion type extraction for three different extraction procedures : 1) ML-Ask basic procedure ; 2) ML-Ask with Shi's technique combined (procedure A); 3) ML-Ask with Shi's technique instead of the usual emotion extraction method (procedure B) (see Fig.2). For each extraction procedure we also compared two variants for both Web mining procedures : 1) extracting and keeping all emotions and 2) taking into account only the emotions that achieved the three highest results, treating the rest as a noise (e.g. in table 5 the three upper types are taken into consideration).

6.3 Applying a 2-dimensional Model of Affect

The idea of a 2-dimensional model of affect was first proposed by Schlosberg (1952) and developed further by Russell (1980). Its main assumption is that all emotions can be described in a space of two-dimensions : the emotion's valence polarity (positive negative) and activation (activated / deactivated). An example of positive-activated emotion would be an "excitement"; a positive-deactivated emotion is, for example, a "relief"; negative-activated and negative-deactivated emotions would be "anger" and "gloom" respectively. This way four areas of emotions are distinguished : activated-positive, activated-negative, deactivated-positive and deactivated-negative (see Figure 4).

The emotions gathered from the Web are then mapped on two-dimensional model of affect and their affiliation to one of the groups is determined. However for some emotion types the affiliation is not obvious (e.g. surprize can be both positive as well as negative; dislike can be either activated or deactivated, etc.) and therefore we mapped them on all of the groups they could belong to. However no emotion type was mapped on more than two adjacent fields. This grouping is then used in the evaluation experiment

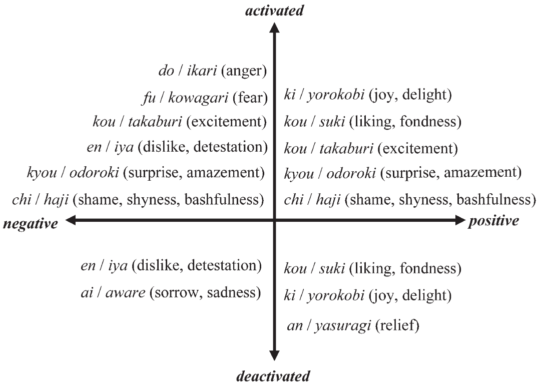

activated
do / ikari (anger)
fu / kowagari (fear)
kou / takaburi (excitement)
en / iya (dislike, detestation)
kyou / odoroki (surprise, amazement)
chi / haji (shame, shyness, bashfulness)
ki / yorokobi (joy, delight)
kou / suki (liking, fondness)
kou / takaburi (excitement)
kyou / odoroki (surprise, amazement)
chi / haji (shame, shyness, bashfulness)
negative
positive
en / iya (dislike, detestation)
ai / aware (sorrow, sadness)
kou / suki (liking, fondness)
ki / yorokobi (joy, delight)
an / yasuragi (relief)
deactivated

Fig.4 Grouping Nakamura's classification of emotions on Russell's two-dimensional space.

for estimating whether the emotive associations gathered from the Web belong to the same space, even if not perfectly matching with the ones tagged by humans (authors of the utterances or the third party evaluators, depending on the phase of evaluation-see below for details).

7. Evaluation- an Inquiry

To verify the system's performance there was a need to carry out an evaluation experiment. Therefore we investigated the methods usually used to evaluate affect analysis systems. Unfortunately, instead of discovering a satisfying method of evaluation, we made an alarming discovery. The majority of the methods we found were insufficient and unobjective.

7.1 Problems with Popular Evaluation Methods

Among many questions that appeared during our research in the field, the most urgent problems with the evaluation methods we noticed in the studies on affect analysis/recognition were as follows.

7.1.1 Searching for Affect in Common Words

Emotions are the domain of human-human communication. Therefore research on recognizing them should be focused on utterances and whole conversations rather than on isolated words. Although there are words which display more emotive coloring than others, the emotiveness of such words becomes visible only in comparison to words similar semantically, but different pragmatically (see examples for Japanese in Table 6).

Table 6 Words with similar semantic-, but different emotive meaning (Kamei, Kouno and Chino 1996)

	Non-emotive	Emotive
Japanese	父	オヤジ
Romanized transcription	*chichi*	*oyaji*
English translation	a father	an old man

However, despite the self-evident nature of the above, there are still scientists developing methods for computing emotiveness of e.g. common nouns, like "pencil" or "laundry" (Saito 2008), which drives us to nonsensical conclusions, like "laundry is joyful and dreadful", and "pencil is favorable and enthusiastic".

7.1.2 Interfering in Evaluation

In creating a corpus for evaluation it is necessary to keep the authors' interference as small as possible. However, it seems popular to interfere in the process of emotional tagging of the corpus (Wu, Chuang and Lin 2006). This might suggest that the tags are set to match the actual results of the system. That would mean that, beginning from the tagging process, ending on the evaluation, the whole work lacks objectivity.

7.1.3 Fallacy of Commonsensical Recognition

Recognition is from the definition a one-target-oriented process. This means that the process of recognition can be performed by at least one recognizer, but only for one target at a time. Even in the case of performing recognition of some features on multiple targets, the process must still be perceived as a multiple number of one-target-oriented processes (Rybak, Golovan and Gusakova 2003). Despite this it is not rare to find works with evaluations of recognition results based on the approximated judgment of a third party of evaluators (Alm, Roth and Sproat 2006 ; Tsuchiya, Yoshimura, Watabe and Kawaoka 2007). This way the very idea of recognition is neglected and comes down to the statement that "if the majority says A, it does not matter what the particular target of the recognition process really thinks-for evaluation puposes they will always think A." When we engage multiple recognizers in the process of evaluation, there are always three kinds of information to be obtained. The first type is, when we assume that the target of recognition is always right, the information on the particular patterns provided by the target of the recognition. The second type of information is the accuracy of guessing the value of those particular patterns by a recognizer/evaluator (or recognizers/evaluators). Finally, the third kind of information is when we neglect the information given by the providers of patterns and check only the general agreement between all the third party evaluators engaged in the process. This standpoint can be called a commonsensical and checks not the pattern or state to be recognized, but the general belief about what is recognized.

To apply this reasoning to the process of recognizing somebody's emotional states, it is crucial to understand that the two kinds of information in question—how other people recognize somebody else's emotional states and what is the general thinking about it—are two completely different kinds of information. Unfortunately, as it was stated above, it is not rare to find research on the so-called commonsense standpoint based recognition.

Both of the information types in question are important for performing an objective evaluation of an affect analysis system. However, basing the evaluation of the process of recognition performed by a machine on general thinking contradicts the very idea of recognition form the very beginning. Although it is true that lenient evaluation methods not taking both of the standpoints into consideration are far easier to perform than creating a fair and reliable database for the evaluation, for example a corpus tagged both by the authors of the utterances and by third party evaluators.

7.1.4 Oversimplified Reasonability

Often the evaluation is oversimplified into asking the evaluators whether the system's results were reasonable (Tsuchiya, Yoshimura, Watabe and Kawaoka, 2007), although Rzepka and Araki state clearly that such ways of evaluation are inappropriate and insufficient since they depend highly on the evaluator's imagination and experiences (Rzepka, Araki 2007).

7.1.5 Small Evaluation

Probably the most common problem, which often goes along with the former ones, is the number of evaluators employed in the evaluation process. If researchers decide to check only how third parties evaluate the results, it is best to ask as many evaluators as possible to get a wide view on the results, count an overall aggreement and rectify potential errors. Unfortunately many scientists limit their evaluation to e.g. five people (Tsuchiya, Yoshimura, Watabe and Kawaoka, 2007), where others even settle for less (Endo, Saito, Yamamoto, 2006). Evaluation limited to such numbers surely provides less cumbersome results, but it is highly questionable whether it is sufficient at all.

7.2 Evaluation of ML-Ask with Popular Evaluation Method

We first used one of the common evaluation methods to evaluate ML-Ask. We assume that the choice of evaluation method greatly influences the view of the results provided by the system, and commonly used simplistic methods could potentially provide positive results regardless of the system's performance. To prove this theory we used Tsuchiya's et al. method which in their opinion decided that the system they proposed was highly effective (accuracy 88%) (Tsuchiya, Yoshimura, Watabe and Kawaoka 2007). The details of this small experiment are described below.

7.2.1 Tsuchiya's Evaluation Method—Short Description

In their evaluation, Tsuchiya et al. asked five people to verify how commonsensical the results given by their system were. The evaluators had three options : A) commonsensical, B) "not uncommonsensical" and C) "uncommonsensical". The result was counted positive for the evaluation if either of the A) or B) options were chosen. Furthermore, if at least two of the five evaluators gave a positive verdict, the system's result was positive. In this rather lenient way Tsuchiya et al. showed that their system achieved 88% accuracy.

7.2.2 ML-Ask in the Perspective of "not uncommonsensicalness"

We performed an evaluation of the results provided by the ML-Ask baseline system using Tsuchiya's method. The system achieved an accuracy of 97%, which would mean that it is almost perfect and simply outperforms all of the present ones. Although it is obviously our goal to achieve that, we would rather achieve such results in an objective evaluation that would give the assurance that our evaluation method was not chosen to increase the results figures.

8. Double Standpoint Evaluation Method (DSEM)

Taking into account the problems stated above, we worked out our own more unbiased method of evaluation for affect analysis systems.

DSEM is a method where the evaluated system is evaluated from two different standpoints : recognitive (the first person evaluation) and commonsensical (the third person evaluation). As a method aiming to be objective, it assumes that neither do people themselves understand their emotional states with 100% reliability, nor do other people perceive the emotional states of their interlocutors with a perfect accuracy. We should rather look for a balance between these two approaches. In our method, the system is first evaluated on whether it can appropriately recognize the emotional states of users. After that, another evaluation is performed by an objective group of third party evaluators to check how much the system's procedures agree with human commonsense about other person's emotions. For both sets of results, 1) the higher the results of both first-and third-person evaluation are, the better, and 2) the more balanced the both results are, the better.

8.1 Corpus for DSEM

For the deep evaluation provided by DSEM there is a need for an appropriate corpus with multi-faceted emotional tagging. As a premise, we do not take isolated words as an object of research. The minimal unit of interest is one whole utterance. By an utterance we mean any act where a set of communicative signs is uttered by a sender to a receiver. It can be simple or consist of a number of sentences.

We have gathered such a corpus. We still continue both : gathering new utterances tagged by their authors, and tagging the corpus gathered this way by a number of third-party human evaluators. Ultimately we plan to gather both short and long utterances and add tags for whole conversation sets in specific contexts.

The evaluation in this paper is based on a corpus of natural utterances gathered through an anonymous survey. In the survey 30 people of different ages and social groups participated. Each of them was to imagine or remember a conversation (or conversations) with any person (or persons) they know and write three sentences from that conversation : one free, one emotive, and one non-emotive. After that the authors tagged the utterances written by themselves in the same way as the system-first whether or not an utterance is emotive. If so, the authors set the emotive value (0-5) and described the specific emotion types conveyed in the emotive utterances.

9. ML-Ask System Evaluation Experiment

We put the proposed method of evaluation into practice to perform a thorough evaluation of the ML-Ask system. We used ML-Ask to analyze our corpus of 90 sentences, and compared the results of the system with the results provided by the authors of the utterances. Although the ideal would be a perfect agreement, it is highly difficult to achieve such a score, since attitudes towards emotive features of speech differ greatly between people. Therefore, we also added an emotive tagging of the corpus by third party human evaluators to determine a general human level in recognizing emotions from text in ordinary people. The third party human evaluators (on average 10 people per sentence) tagged the corpus emotively in the same way as the system does and for each of them we counted the accuracy in recognition by comparing the results to the results provided by the authors of the utterances. The level of results achieved by the third party evaluators was considered to be the human level of accuracy in recognizing emotions from text. This provides us with the evaluation from the first person (recognitive) perspective.

However, as it was pointed out before, it cannot be taken as a certainty that the authors of the utterances know perfectly the state of emotions they are conveying and the desired result is the one which is high and balanced for both first person and third person perspective evaluation. Therefore, to broaden the evaluation, the second (commonsense) perspective is also applied, by having the same utterances tagged by third party human evaluators. Our mehtod assumes the higher relevance the more evaluators take part in the third party evaluation. In this paper 10 people on average evaluated every utterance. Although we agree that it is not enough, it is still two - to ten times bigger then the number of evaluators used in other research. After gathering the results of the third party evaluators we calculate an average agreement between them, which we consider as a general commonsense. After that we calculate the system's agreement with the commonsense to check the commonsense level of the evaluated method. In our assumption, the more similar the system was to the average impressions of the people, the closer it was to the level of human commonsense.

9.1 Recognition Accuracy Evaluation

To calculate the system's true performance in recognition we evaluated three levels of analysis performed by the system : 1) the analysis of general emotiveness ; 2) calculating the emotive value ; and 3) determining the specific emotion types. In this phase of evaluation we use the taggings made by the authors of the utterances as a base. Then we checked the level of human accuracy in recognition by comparing base taggings to the taggings made by third party evaluators. The system's closeness to this level is also the percentage of how high the system's performance is comparable to humans.

9.1.1 Emotive / non-emotive

To calculate the system's true performance or total accuracy in determining the general emotiveness, we calculated the balanced F-score for finding emotive and non-emotive utterances separately.

The former F_E (equation 1) is calculated on the basis of precision P_E and recall R_E. P_E is calculated as the number of the utterances correctly judged by the system as emotive S_{CE} divided by all utterances judged to be emotive S_{DE} (equation 2). R_E is calculated as the number of utterances correctly judged by the system as emotive S_{CE} dvided by all emotive utterances S_{AE} (equation 3).

$$F_E = \frac{2 \cdot (P_E \cdot R_E)}{(P_E + R_E)} \quad (1)$$

$$P_E = \frac{S_{CE}}{S_{DE}} \quad (2)$$

$$R_E = \frac{S_{CE}}{S_{AE}} \quad (3)$$

Then F_{NE} (equation 4) is calculated on the basis of precision P_{NE} and recall R_{NE}. P_{NE} is calculated as the number of the utterances correctly judged by the system as non-emotive S_{CNE} dided by all utterances judged to be non-emotive S_{DNE} (equation 5). R_{NE} is calculated as the number of the utterances correctly judged by the system as non-emotive S_{CNE} divided by all utterances judged as non-emotive S_{ANE} (equation 6).

$$F_{NE} = \frac{2 \cdot (P_{NE} \cdot R_{NE})}{(P_{NE} + R_{NE})} \quad (4)$$

$$P_{NE} = \frac{S_{CNE}}{S_{DNE}} \quad (5)$$

$$R_{NE} = \frac{S_{CNE}}{S_{ANE}} \quad (6)$$

The total accuracy of the system in determining whether an utterance is emotive or not is the average F-score $F_{E/NE}$ for recognizing both emotive F_E (equation 1) and non-emotive F_{NE} (equation 4) utterances and is calculated as in equation (7).

$$F_{E/NE} = \frac{F_E + F_{NE}}{2} \quad (7)$$

The total accuracy of ML-Ask in determining the general emotivenes evaluated on 90 items was $F_{E/NE} = 0.83$. The result is very promising, since the same value for human evaluators had a wide range of results from 0.4 to 0.86. Since the result of ML-Ask was placed close to the top of this ranking, we can say that the system recognizes emotiveness on a very high level, comparable to humans.

9.1.2 Emotive value

Determining the emotive value of an utterance is highly dependent on many constantly changing situational features of a conversation as well as on the personal experiences of the interlocutors. Therefore it is close to impossible to achieve a perfect match for all of the utterances with analysis of only the textual layer of an utterance. To account for this, we used a more lenient condition of almost-perfect match (when the emotive value differs between speaker and system by ±1 emotive point for an utterance) in the process of evaluating the system's agreement with the speaker.

The accuracy of determining the emotive value with this condition for the available 30 items tagged as described in section 8.2, an agreement coefficient[3] of 0.5 was achieved. As for the third party human evaluators, at the time of the experiment there were only two small sets of utterances tagged this way. In one with 19 utterances, 16 were tagged correctly. In the second one, for 12 utterances 10 were tagged correctly. This shows a tendency of 0.84 of agreement coefficient for recognition accuracy of emotive value by humans. The system's result is 60% of the human level. It is thus desirable to find a more precise way of setting the emotive value in the future than the quantitative one used at this stage.

9.1.3 Specific Emotion Types

By recognizing a specific type of emotion we mean correctly recognizing any but at least one emotion assigned by the author for the utterance, including non-emotive. However, people often misassign their specific types of emotions, but usually guess whether the emotion type they have is positive or negative, and whether it is activated (aroused), or deactivated (forceless). Therefore additionally, the result was considered as positive also when the extracted emotions were belonging to the same quarter of Russell's 2-dimmensional space, even if the extracted emotions were not exactly the same as the ones annotated by the authors of the utterance. The emotion type recognition accuracy is calculated as F_{ETR} (equation 8) value from the approximation of the accuracy to determine about non-emotiveness F_{NE} (see 9.1.1), and the accuracy to determine about the specific emotion types F_{ET}.

$$F_{ETR} = \frac{F_{ET} + F_{NE}}{2} \tag{8}$$

F_{ET} is calculated based on the precision P_{ET} and the recall R_{ET} values (equation 9). P_{ET} is calculted as the number of utterances with correctly judged emotion types S_{CET} divided by all utterances with described emotion types S_{DET} (equation 10). R_{ET} is calculated as the number of utterances with correctly described emotion types S_{CET} divided by all utterances with particular emotion types set by the authors S_{AET} (equation 11). There were 30 items available for the evaluation.

$$F_{ET} = \frac{2 \cdot (P_{ET} \cdot R_{ET})}{(P_{ET} + R_{ET})} \tag{9}$$

$$P_{ET} = \frac{S_{CET}}{S_{DET}} \tag{10}$$

$$R_{ET} = \frac{S_{CET}}{S_{AET}} \tag{11}$$

Basic Procedure

The system's result in estimating the specific types of emotions was a balanced F-score of 0.45. As for human evaluators, the average result was 0.72. Therefore the system's accuracy was approximately 63% (0.45/0.72) of the human level. In future research it is desirable to both improve the algorithm for determining the specific emotion types and to perform the evaluation on a larger number of tagged utterances. We also calculated times for all of the procedures to check whether the system is capable to operate in real time. The approximate time of processing one utterance in the basic procedure was 0.143 s.

Procedure A

This method for emotion type extraction used a Web mining technique as a support for ML-Ask in all cases where the system could determine that an utterance was emotive but the basic procedure was not able to determine the emotion type. ML-Ask supported with this method achieved a balanced F-score of 0.54 and 0.53 for the two variants "3 best results" and "all emotions" respectively. This gives approximately 75% (0.54/0.72) and 74% (0.53/0.72) of the human level in determining the emotion types.

Procedure B

In this method for emotion type extraction, the Web mining technique is always used instead of the system's lexical extraction (basic procedure).

[3] For a detailed description of the agreement coefficient see 9.2

ML-Ask with this method achieved a balanced F-score of 0.61 and 0.53 respectively of for the two variants. Compared to the human level in determining the types of emotions this amounts to 85% (0.61/0.72) and 74% (0.53/0.72) respectively. The approximate time of processing one utterance in both procedures using Web mining, regardless of the length of the utterance, was from a minute to half an hour. The time lags were caused by problems with the network and, in many cases, by errors of the search engine API.

9.2 Commonsense evaluation

In the evaluation from the commonsensical standpoint, the system's agreement with the generally understood commonsense is evaluated. We perform the evaluation by checking how many of the results provided by the system coincide with the results of the third party human evaluators. In this stage we do not take into consideration the authors' tags of the utterances. Also, since in this process we check the agreement of one machine-evaluator (ML-Ask) with all the human evaluators, it is difficult to use terms like "precision" to calculate the system's true accuracy. The F-score is thus inappropriate for this stage of evaluation. Therefore we calculate a coefficient of agreement in the results between the system and the third party evaluators.

The agreement *U* between one evaluator Eva_A and another Eva_B is the ratio of similarly tagged utterances (t_{sim}) to the total number of utterances (t_i) tagged by the human evaluators as showed in equation (12). The agreement coefficient is counted for every pair of the human evaluators. Then, the average agreement of one evaluator with the rest is calculated for each of the evaluators. The average gives us the level of agreement of one evaluator with all the rest - the equivalent of commonsense. After that the system's agreement with each of the humans is calculated. If system' s agreement coefficient fits into the scope of agreement between humans - the system's results are judged as commonsensical.

$$U_{Eva_B}^{Eva_A} = \frac{t_{sim}}{t_i} \qquad (12)$$

9.2.1 Emotive / Non-emotive

The agreement of each human evaluator with all other human evaluators (there were eight human evaluators ; 2 females and 6 males) in determining whether the utterance is emotive or not for 60 items aviable for this stage of evaluation gave a very wide range, from 0.62 to 0.95. However, the average agreement between all human evaluators was from 0.70 to 0.82 (see Table 7). The level of human commonsense in determining the emotiveness in ML-Ask was the system's distance from to the average agreements between human evaluators.

The system's average agreement with all evaluators

Table 7 Results of agreement in emotiveness determination between the evaluators and between the system and the rest of the evaluators (the two of the most extreme results are highlighted in boldface ; for average values the two highest and two lowest)

	ML-Ask	Eva01	Eva02	Eva03	Eva04	Eva05	Eva06	Eva07	Eva08
ML-Ask		**0.78**	0.6	0.5	0.55	0.58	**0.47**	0.66	0.53
Eva01	**0.78**		0.75	0.68	0.68	0.73	**0.62**	0.76	0.68
Eva02	0.6	0.75		0.8	0.78	0.85	0.87	0.78	0.9
Eva03	0.5	0.68	0.8		0.71	0.88	0.83	0.81	0.9
Eva04	0.55	0.68	0.78	0.71		0.68	0.71	0.7	0.68
Eva05	0.58	0.73	0.85	0.88	0.68		0.82	0.79	**0.95**
Eva06	**0.47**	**0.62**	0.87	0.83	0.71	0.82		0.74	0.87
Eva07	0.66	0.76	0.78	0.81	0.7	0.79	0.74		0.78
Eva08	0.53	0.68	0.9	0.9	0.68	**0.95**	0.87	0.78	
Average (approx.)	0.58	**0.7**	**0.82**	0.8	**0.7**	**0.81**	0.78	0.76	**0.82**
The highest and the lowest agreement between the system and humans	0.47– 0.78	Bracket for human level of commonsense in determining emotiveness						0.7 – 0.82	
		The highest and the lowest agreement between humans						0.62 – 0.95	

reached 0.58, which is above 83% of the human level of commonsense. The result is lower than the result from the recognitive standpoint evaluation, where the system reached human level (0.83 of the balanced F-score). The difference in results shows how a different standpoint can change the view on the results. The result however is still very promising, since improving the recognition will likely also improve the commonsense level. This also indicates that the direction of development of the system is correct.

As an interesting fact we might add that for sentences with a perfect match, where both authors and all evaluators were unanimous about the emotiveness (18 of 60 sentences), the system's result was 100%.

9.2.2 Emotive Value

With the condition of almost-perfect match (see 8.1.2), the agreement of ML-Ask with the eight human evaluators (2 females and 6 males) was in the range from 0.4 to 0.57 (average=0.51). The intervals of the average agreements between human evaluators was 0.61 to 0.75. With the average agreement being 0.51 the system achieved 84% of the human level of commonsense in determining the emotive value. This result confirms the result from the emotiveness evaluation (83%). Although it is desirable to improve the way of determining the emotive value for the system to bahave more similarly to human commonsense, we consider this result as promising, since the emotional intensity setting is highly subjective and ambiguous among people. It is also likely that improving the accuracy in determining the general emotivenes of a sentence will positively influence the accuracy of the emotive value determination.

We also made some interesting observations. In both commonsense evaluations-of emotiveness and of emotive value-similar tendencies in levels of agreement with other evaluators can be seen clearly (compare Tables 7 and 8). This indicates that the evaluation method is reasonable and effective.

9.2.3 Specific Emotion Types

For evaluation of specific emotion judgements by the system compared to the general commonsense we performed another survey. We asked twelve different evaluators (2 females, 10 males) about emotions conveyed in emotive utterances (with a possibility of specifying more than one emotion). In many cases the results differed significantly and there were sentences in which evaluators were not able to identify any emotions. With this in mind the following allowances were made. If ML-Ask extracted from a sentence at least one of the emotion types classified by evaluators (see examples in table 10) or the system's classification coincided with the majority (the condition applicable mostly when tha majority of the evaluators judged no emotion types), the result was positive.

Table 8 Results of agreement in emotive value determination between the evaluators one another and between the system and the rest of the evaluators (the two of the most extreme results are highlighted in boldface ; for average values the two highest and two lowest)

	ML-Ask	Eva01	Eva02	Eva03	Eva04	Eva05	Eva06	Eva07	Eva08
ML-Ask		0.57	0.55	**0.4**	0.52	0.53	**0.5**	0.55	0.43
Eva01	**0.57**		0.7	0.67	**0.53**	0.8	0.57	0.62	0.6
Eva02	0.55	0.7		0.75	0.68	0.82	0.85	0.73	0.77
Eva03	**0.4**	0.67	0.75		0.57	0.72	0.65	0.57	0.73
Eva04	0.52	**0.53**	0.68	0.57		0.68	0.63	0.58	0.6
Eva05	0.53	0.8	0.82	0.72	0.68		0.75	0.72	**0.95**
Eva06	0.5	0.57	0.85	0.65	0.63	0.75		0.75	0.83
Eva07	0.55	0.62	0.73	0.57	0.58	0.72	0.75		0.75
Eva08	0.43	0.6	0.77	0.73	0.6	**0.95**	0.83	0.75	
Average(approx.)	0.51	**0.64**	**0.76**	0.67	**0.61**	**0.78**	0.72	0.67	**0.75**
The highest and the lowest agreement between the system and humans	0.4– 0.57	Bracket for human level of commonsense in determining emotive value						0.61 – 0.75	
		The highest and the lowest agreement between humans						0.53 – 0.95	

Basic procedure

In the commonsense evaluation the basic method for emotion extraction reached 45% of the human commonsense level. Three examples of successful outputs are shown in Table 10.

The result is not perfect, and in the future we want to improve the accuracy in determining the particular emotion types. However, the result is still encouraging, since its imperfection arises from, for instance, lack of appropriate and up-to-date entries in databases created from Nakamura's collection (Nakamura 1993). However, we have already started to retrieve new entries from the Internet by using keywords from his collection to update the database, which will most probably lead to improvements. Also improving the accuracy in both determining general emotiveness and emotive value will likely to rise the commonsense level.

Procedure A

The commonsense level of this method was evaluated to 67% and 60% respectively.

Procedure B

The commonsense level of this method was evaluated to 60% and 50% respectively.

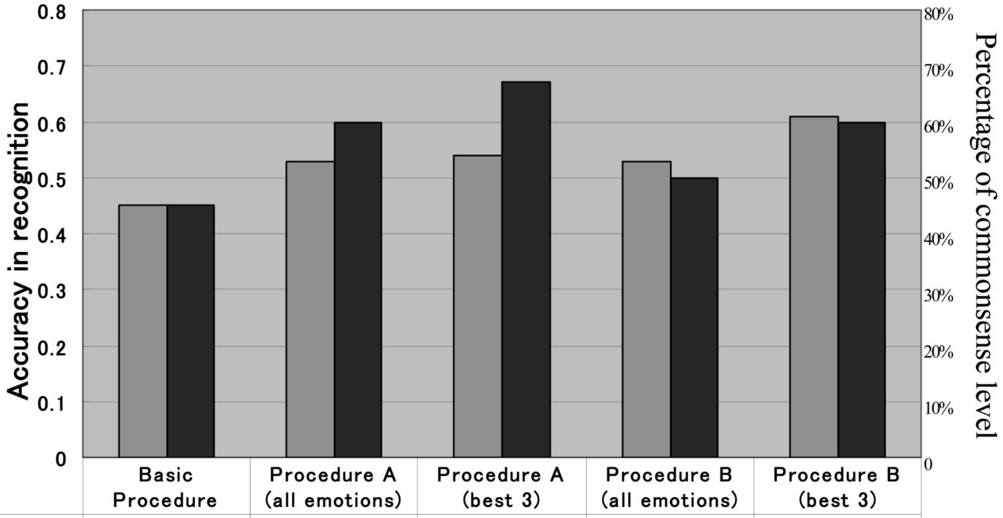
Two examples of successful outputs for each method are shown in Table 9. All the results are summarized on Figure 5.

Table 9 Two examples of improvements over the basic procedure with the Web mining technique in recognizing emotion types

Utterance in Japanese / ***romanized reading /*** **meaning in English**	**Author of the utterance (first person tagging)**	**ML-Ask basic procedure**	**ML-Ask with Web mining (only "3-best" variant; lower example – procedure A, upper example – procedure B)**	**Human evaluators (third person tagging)**
諦めちゃいけないよ / *Akiramecha ikenai yo* / Don't give up !	昂 [excitement], 喜 [joy]	厭 [dislike, aversion]	喜 [joy], 哀 [sadness], 昂 [excitement], 驚 [surprise]	昂 [excitement], 好 [liking, fondness], 怒 [anger], 厭 [dislike, aversion]
映画の最後の部分で、思ってもいなかった反応をしちゃったんですよ！！！/ *Eiga no saigo no bubun de, omotte mo ina katta hannou wo shichatta n desu yo!!!* / Watching the last scenę of the movie I reacted surprisingly even for myself!!!	驚 [surprise]	無 [nothing]	喜 [joy], 驚 [surprise], 好 [liking, fondness]	昂 [excitement], 喜 [joy]

Table 10 Three examples of successful recognition of emotion types in the commonsense evaluation of the basic procedure

Utterance in Japanese / ***romanized reading /*** **meaning in English**	**ML-Ask (basic procedure)**	**Human evaluators**
あなたの事が好きなんです。/ *Anata no koto ga suki nan desu.* / I love you.	好 [liking, fondness]	好 [liking, fondness] (7), 怖 [fear] (3), 喜 [joy] (2), 恥 [shame, shyness, bashfulness] (1), 昂 [excitement] (1)
この本さー、すげーやばかったよ。まじ怖すぎ。 / *Kono hon saa, sugee yabakatta yo. Maji kowa sugi.* / That book, ya know, 't was a killer. It was just too scary.	怖 [fear]	怖 [fear] (6), 驚 [surprise] (2), 昂 [excitement] (2), 喜 [joy] (1), 無 [nothing] (1)
うん、うまい、感激だ。/ *Un, umai, kangeki da.*/ Yeah, it's great, I'm impressed.	昂 [excitement]	喜 [joy] (10), 昂 [excitement] (5), 驚 [surprise] (1), 好 [liking, fondness] (1)



	Basic Procedure	Procedure A (all emotions)	Procedure A (best 3)	Procedure B (all emotions)	Procedure B (best 3)
■ Recognition	0.45	0.53	0.54	0.53	0.61
■ Commonsense	45%	60%	67%	50%	60%

Fig.5 Summarized results for 3 different procedures of emotion types eztraction with 2 variants of procedures using Web mining technique

10. Conclusions and Future Work

In this paper we presented a method for affect analysis of utterances, supported with a Web mining technique.

The recognizing emotiveness from textual input experiment confirmed that emotiveness is not incomputable on the lexical level. Moreover, a computer system designed in a specified way can determine emotiveness of a sentence with human level accuracy, within specific limits described in this paper. Although there were still problems not yet solvable for a machine, the general results were very encouraging and thus we will continue the works on this project.

In this paper we also presented DSEM-an objective method suitable for evaluating systems analyzing and recognizing emotions. The method is based on two standpoints of evaluation—recognitive and commonsensical. We put the method into practice to evaluate the ML-Ask system. The evaluation gave very promising results, but also helped us to realize what should be a question of concern in the future research on the project.

To verify whether the evaluation method proposed by us actually gives wider and more objective view on the results, we compared DSEM to the most popular method in the field today. In comparison DSEM clearly revealed drawbacks of the latter. In the most popular method of evaluation used in the comparison, the results of the system are shown to only a few human evaluators, who determine whether the results are commonsensical or not. With this method ML-Ask achieved almost perfect accuracy. On the other hand, DSEM, although requiring more effort to perform, shows the results more accurately, without distortions and bending, and provides clear information on what parts of the system should be improved. It would be good if this method was accepted widely in the field.

ML-Ask achieved a high accuracy result of 0.83 balanced F-score in recognizing the general emotiveness of an utterance. This level was confirmed in the commonsensical evaluation, with achieving a close result, above 83% of the general commonsense level.

The emotive value of an utterance was recognized by the system with an agreement coefficient of 0.5, which is 60% comparing to human level of recognition. Although the method was objectively confirmed as close to commonsensical (84%), we should improve the method of determining the emotive value to be closer to the speakers' intention.

The system recognizes specific types of emotions conveyed in utterances on a fair but improvable level of 0.45 balanced F-score. This level was also confirmed the commonsense evaluation (45%).

In every case the Web mining improved the accuracy of the system in extracting the specific types of emotions, although there were some differences in commonsense levels found for the two extraction procedures we evaluated in this paper. There is a need for further experiments on a larger evaluation material to determine which is the most accurate extraction procedure. Experiments on different variants of the method showed that it is more effective to keep only the emotions that achieved the three best results and discard the rest as a noise, rather than keeping all of the extracted emotions. Furthermore, since the system's procedures are used sequentially, improving accuracy in earlier stages is very likely to improve the emotion type extraction. The tools for morphological analysis used in the system are also not perfect, which might decrease the system's real accuracy. However improving the extraction of n-gram phrases from the sentences to analyze should eliminate these difficulties. In the near future we also plan to apply Russell's (1980) two-dimensional model of emotions to reduce the ambiguities in the database of emotive elements, which should also enhance the emotion type extraction.

The baseline system can be used in real-time applications, since the approximate time for processing one utterance is 0.143s. The processing times in the procedures using Web mining are rather long (several minutes or more), which might be caused by potential problems with network connections or imperfections in search engine indexing. However, this can be mitigated by using a different search engine, or applying a Web page indexing method created especially for the needs of this research, for example using HyperEstraier, which we plan to do in the near future.

Acknowledgments

This research is partially supported by a Research Grant from the Nissan Science Foundation and The Global Centers of Excellence Program founded by Japan's Ministry of Education, Culture, Sports, Science and Technology.

References

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

Alm, Cecilia Ovesdotter, Roth, Dan and Sproat, Richard. "Emotions from text : machine learning for text based emotion prediction." *Proceedings of HLT/EMNLP*, pp. 579 - 586, Vancouver. 2005.

Argyle, Michael. *The psychology of interpersonal behavior*, Penguin, Baltimore. 1967.

Baba, Junko. "Pragmatic Functions of Japanese Mimesis in Emotive Discourse." Southern Japan Seminar, Atlanta. 2001

Baba, Junko. "Pragmatic function of Japanese mimetics in the spoken discourse of varying emotive intensity levels." *Journal of Pragmatics*, vol. 35, no.12, pp. 1861 - 1889, Elsevier. 2003

Beijer, Fabian. "The syntax and pragmatics of exclamations and other expressive/emotional utterances." *Working Papers in Linguistics 2*, The Department of English in Lund. 2002

Bong-Seok Kang, Chul-Hee Han, Sang-Tae Lee, Dae-Hee Youn and Chungyong Lee. "Speaker dependent emotion recognition using speech signals." In *Proc. ICSLP*, pp. 383 - 386. 2000.

Bühler, Karl. *Sprachtheorie. Die Darstellungsfunktion der Sprache.* (Theory of Language. Representative Function of Language.), Ullstein, Frankfurt a. M., Berlin, Wien. 1978

Crystal, David. *The Cambridge Encyclopedia of Language.* Cambridge University Press. 1989

Darwin, Charles R. *The Expression of the Emotions in Man and Animals*, John Murray, London. 1872

Dybala, Pawel, Ptaszynski, Michal, Rzepka, Rafal and Araki, Kenji. "Extracting Dajare Candidates from the Web - Japanese Puns Generating System as a Part of Humor Processing Research." *Proceedings of the First International Workshop on Laughter in Interaction and Body Movement (LIBM'08)*. 2008

Dybala, Pawel, Rzepka, Rafal and Araki, Kenji. "Dajare Generating Support Tool - Towards Applicable Linguistic Humor Processing." *Proceedings of The Fourteenth Annual Meeting of The Association for Natural Language Processing*, pp.701 - 704. 2008

Dybala, Pawel, Rzepka, Rafal and Araki, Kenji "Dajare Types and Individualized Sense of Humor - A Prelude to PUNDA Project." *Proceedings of 2007 Joint Convention Record*, *The Hokkaido Chapters of the Institutes of Electrical and Information Engineers*, Japan, pp. 293 - 294. 2007a

Dybala, Pawel, Rzepka, Rafal and Araki, Kenji. "PUNDA Project - a Design For A Japanese Puns Generating system." *Language Acquisition and Understanding (LAU) Technical Report*, Hokkaido University, pp. 6 - 11. 2007b

Endo, D., Saito, S. and Yamamoto, K. "Kakariuke kankei wo riyo shita kanjoseikihyogen no chushutsu." (Extracting expressions evoking emotions using dependency structure), *Proceedings of The Twelve Annual Meeting of The Association for Natural Language Processing.* 2006

Frijda, Nico H. *The Emotions*. Cambridge University Press, Cambridge. 1987

Geiser, Georg. *Mensch-Maschine Kommunikation*. Oldenbourg. 1990.

Hall, Edward T. *The Hidden Dimension*. Double Day, New York. 1966.

Hase, Masao, Shiori Kenta, and Hoshino Junichi. "Hatsuwa wo okonau kagu ni yoru nichijouteki entateinmento (taiwa) [The Everyday Entertainment by Talking Furniture]" (in Japanese). *IPSJ SIG Technical Report* 2007-NL-181, 2007 (94), pp.41-46. 2007.

Higuchi, Shinsuke, Rzepka, Rafal and Araki, Kenji. "A Casual Conversation System Using Modality and Word Associations Retrieved from the Web." *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.382-390, Honolulu, USA, 2008

Hirabayashi, Mikio, "HyperEstraier", (http://hyperestraier.sourceforge.net/index.html), 2004-2007.

Itoh, Toshihiko, Minematsu, Nobuaki and Nakagawa, Seiichi. "Analysis of filled pauses and their use in a dialogue system." *The Journal of the Acoustical Society of Japan*, 55 (5), pp. 333-342. 1999.

Izard, Carroll Ellis. *Human Emotions*. Springer. 2007.

Jakobson, Roman. "Closing Statement : Linguistics and Poetics." *Style in Language*, The MIT Press, Massachusetts, pp. 350-377, 1960.

Kamei, Takashi, Kouno, Rokuro and Chino, Eiichi. *The Sanseido Encyclopedia of Linguistics*, Vol.6, Sanseido. 1996

Kudo, Taku. "MeCab : Yet Another Part-of-Speech and Morphological Analyzer." 2001,
http://mecab.sourceforge.net/ (2008. 06. 01).

Matsumoto, Kazuyuki, Ren, Fuji, Kuroiwa, Shingo, Tsuchiya Seiji."Emotion Estimation Algorithm Based on Interpersonal Emotion Included in Emotional Dialogue Sentences", *Lecture Notes in Computer Science 4827*, pp.1035-1045, Berlin, Springer-Verlag, 2007.

Matsumoto, Yuji, Kitauchi, Akira, Yamashita, Tatsuo, Hirano, Yoshitaka, Matsuda, Hiroshi, Takaoka, Kazuma and Asahara, Masayuki. "Japanese Morphological Analysis System ChaSen version 2.2.1." Nara Institute of Science and Technology. 2000

Morris, Merrill and Ogan Christine. "The Internet as Mass Medium." *Journal of Computer-Mediated Communication*, 1 (4). 1996.

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

Nakayama, N., Eguchi, K. and Kando, N. "A Proposal for Extraction of Emotional Expression." *IEICE Technical Report*, vol. 104, no.416, pp.13-18. 2004

Ono, Hajime. "An emphatic particle DA and exclamatory sentences in Japanese." University of California, Irvine. 2002.

Ortony, Andrew, Clore, Gerald L. and Collins, Allan. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge. 1988

Oshima-Takane, Y., MacWhinney B. (Ed.), Shirai, H., Miyata, S. and Naka, N.(Rev.) *CHILDES Manual for Japanese*, McGill University, The JCHAT Project. 1995-1998

Picard, Rosalind W. "Affective Computing." *MIT Technical Report* #321, MIT Media Laboratory. 1995

Picard, Rosalind W. *Affective Computing*, The MIT Press, Cambridge. 1997.

Potts, Christopher and Kawahara, Shigeto. "Japanese honorifics as emotive definite descriptions."*Proceedings of Semantics and Linguistic Theory*, 14, pp.235-254. 2004.

Ptaszynski, Michal. "Moeru gengo-Intānetto kei-jiban no ue no nihongo kaiwa ni okeru kanjōhyōgen no kōzō to kigōrontekikinō no bunseki- "2channeru" denshikeijiban o rei toshite-" (Boisterous language. Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum - 2channel -) , M.A. Dissertation, UAM, Poznan. 2006

Ptaszynski, Michal. and Sayama, Kohichi. "The idea of dynamic memory management system based on a forgetting-recalling algorithm with emotive analysis." *Language Acquisition and Understanding (LAU) Technical Report*, Hokkaido University, pp.12-16. 2007

Ptaszynski, Michal, Dybala, Pawel, Rzepka, Rafal and Araki, Kenji. "Effective Analysis of Emotiveness in Utterances Based on Features of Lexical and Non-Lexical Layers of Speech." *Proceedings of The Fourteenth Annual Meeting of The Association for Natural Language Processing*, pp.171-174. 2008a

Ptaszynski, Michal, Dybala, Pawel, Rzepka, Rafal and Araki, Kenji. "Double Standpoint Evaluation Method for Affect Analysis System." *Proceedings of JSAI2008*. 2008b.

Ptaszynski, Michal, Dybala, Pawel, Shi, Wenhan, Rzepka, Rafal and Araki, Kenji."Lexical Analysis of Emotiveness in Utterances for Automatic Joke Generation", *ITE Technical Report* Vol.32, No.47, pp.39-42, The Institute of Image Information and Television Engineers. 2007

Ptaszynski, Michal, Sayama, Kohichi, Rzepka, Rafal and Araki, Kenji. "A dynamic memory management system based on forgetting and recalling." *Proceedings of 2007 Joint Convention Record, The Hokkaido Chapters of the Institutes of Electrical and Information Engineers*, Japan, pp. 295-296. 2007

Read, Jonathon. "Recognizing Affect in Text using Pointwise-Mutual Information." M. Sc. Dissertation, University of Sussex. 2004

Rybak, Ilya A., Golovan, Alexander V., Gusakova, Valentina I. "Behavioral model of visual perception and recognition." *Proceedings of SPIE*, Vol. 1913, pp. 548-560. 1993

Russell, James A. "A circumplex model of affect." *Journal of Personality and Social Psychology*, 39 (6), pp.1161-1178. 1980.

Rzepka, Rafal and Araki, Kenji "What About Tests In Smart Environments? On Possible Problems With Common Sense In Ambient Intelligence." *Proceedings of 2nd*

Workshop on Artificial Intelligence Techniques for Ambient Intelligence, *IJCAI'07*. 2007

Rzepka, Rafal, Ge, Yali and Araki, Kenji. "Common Sense from the Web? Naturalness of Everyday Knowledge Retrieved from WWW." *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 10 (6), pp.868－875. 2006.

Saito, T., Mitsube Y., Nakaya K., Han D., "Meishi no kanjoshozokusei no chushutsu to sore ni motozuku meishi no ruijido no keisan" (Computing similarity of nouns on the basis of noun emotional affiliation), *Proceedings of The Fourteenth Annual Meeting of The Association for Natural Language Processing*, pp.368－371. 2008

Schlosberg, H. "The description of facial expressions in terms of two dimensions." *Journal of Experimental Psychology*, 44, pp.229－237. 1952.

Schwarz, Norbert. "Emotion, cognition, and decision making." *Cognition & Emotion*, 14 (4), pp.433－440. 2000.

Shi, Wenhan. "Emotive Information Discovery from User Textual Input Using Causal Associations from the Internet." *FIT2008*, pp.267－268, 2008.

Singhal, Amit. "Modern Information Retrieval : A Brief Overview." *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24 (4), pp.35－43. 2001.

Sjöbergh, Jonas. "Vulgarities are fucking funny, or at least make things a little bit funnier." *Technical Report of KTH*, Stockholm. 2006

Sjöbergh, Jonas and Araki, Kenji. "A Multi－Lingual Dictionary of Dirty Words", LREC, 2008.

Stevenson, Charles L. *Facts and Values－Studies in Ethical Analysis*. Yale University Press. 1963.

Takahashi, Toshiaki, Watanabe, Hiroshi, Sunda, Takashi, Inoue, Hirofumi, Tanaka, Ken'ichi and Sakata, Masao. *Technologies for enhancement of operation efficiency in 2003i IT Cockpit*, Nissan Technical Review, 53, pp.61－64. 2003.

Tokuhisa, Ryoko, Inui, Kentaro, Matsumoto, Yuji. "Emotion Classification Using Massive Examples Extracted from the Web", *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp.881－888, Manchester, August 2008

Tsuchiya, Naoko. "Taiwa ni okeru kandōshi, iiyodomi no tōgoteki seishitsu ni tsuite no kōsatsu. [Statistical observations of interjections and faltering in discourse]" (in Japanese), *SIG－SLUD－9903－11*. 1999.

Tsuchiya, Seiji, Yoshimura, Eriko, Watabe, Hirokazu and Kawaoka, Tsukasa. "The Method of the Emotion Judgement Based on an Association Mechanism." *Journal of Natural Language Processing*, Vol.14, No.3, The Association for Natural Language Processing. 2007

Wu, Chung－Hsien, Chuang, Ze－Jing, and Lin, Yu－Chung. "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models." *ACM Transactions on Asian Language Information Processing*, 5 (2), pp.165－183. 2006.

Yoshitaka Yamashita. "Kara, Node, Te-Conjunctions which express cause or reason in Japanese (in Japanese)." *Journal of the International Student Center*. Hokkiado University, 3. 1999.

Young, Polly. "The Effects Of Mood And Emotional State On Decision Making." *The Journal of Behavioral Decision Making : Special Issue*. 2006.

（2008年８月15日　受付）
（2008年12月22日　採録）

[Contact Address]
Kita－ku, Kita 14 Nishi 9, 060－0814 Sapporo, Japan
Graduate School of Information Science and Technology, Hokkaido University
Michal Ptaszynski
TEL：011-706-7389(7389)
FAX：011-709-6277
E-mail：ptaszynski@media.eng.hokudai.ac.jp

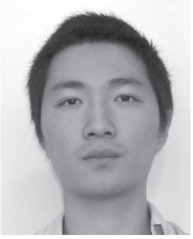
Information about Author

Michal PTASZYNSKI [member]

Michal Ptaszynski was born in Wroclaw, Poland in 1981. He received his M.A. from the University of Adam Mickiewicz in Poznan, Poland in 2006. He was a research student at Otaru University of Commerce, and since 2007 he is studying towards his Ph.D. degree at Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interest includes Natural Language Processing, Dialogue Processing, Affect Analysis, Sentiment Analysis and Information Retrieval. He is a member of the SOFT, JSAI and NLP.

Pawel DYBALA [non-member]

Pawel Dybala was born in Ostrow Wielkopolski, Poland in 1981. He received his M.A. from the Jagiellonian University in Krakow, Poland in 2006. He was a research student at Hokkaido University, and since 2007 he is studying towards his Ph.D. degree at Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interest includes Natural Language Processing, Dialogue Processing, Humor Processing, and Information Retrieval.



Wenhan SHI [non-member]

Wenhan Shi was born in Shenyang, People's Republic of China, in 1985. He received his B.E. from Hokkaido University, Japan. Since 2008 he is studying towards his M.A. degree at Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interest includes Natural Language Processing, Web Mining, Affect Analysis, and Information Retrieval.

Rafal RZEPKA [non-member]

Rafal Rzepka was born in Szczecin, Poland in 1974. He received his M.A. from the University of Adam Mickiewicz in Poznan, Poland in 1999 and Ph.D. from Hokkaido University, Japan in 2004. Now he is an assistant professor at Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interest includes natural language processing, web mining, commonsense retrieval, dialogue processing and affect analysis. He is a member of the AAAI, ACL, JSAI, IPSJ, IEICE, JCSS and NLP.

Kenji ARAKI [non-member]

Kenji Araki was born in 1959 in Otaru, Japan. He received B.E., M.E. and Ph.D. degrees in electronics engineering from Hokkaido University, Sapporo, Japan in 1982, 1985 and 1988, respectively. In April 1988, he joined Hokkai Gakuen University, Sapporo, Japan. He was a professor of Hokkai Gakuen University. He joined Hokkaido University in 1998 as an associate professor of the Division of Electronics and Information Engineering. He was a professor of the Division of Electronics and Information Engineering of Hokkaido University from 2002. Now he is a professor of the Division of Media and Network Technologies of Hokkaido University. His interest is natural language processing, spoken dialogue processing, machine translation and language acquisition. He is a member of the AAAI, IEEE, JSAI, IPSJ, IEICE and JCSS.