知能と情報（日本知能情報ファジィ学会誌）Vol.22, No.1, pp.73－89（2010）

原著論文

An Automatic Evaluation Method for Conversational Agents Based on Affect－as－Information Theory

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

This paper presents a method for automatic evaluation of conversational agents. The method consists of several steps. First, an affect analysis system is used to detect users' general emotional engagement in the conversation and classify their specific emotional states. Next, we interpret this data with the use of reasoning based on Affect-as-Information Theory to obtain information about users' general attitudes to the conversational agent and its performance. The affect analysis system was also enhanced with a procedure for analysis of Contextual Valence Shifters to help determine the semantic orientation of emotive expressions. The method is used as a background procedure during users' conversations with two Japanese－speaking conversational agents. To verify the usability of the method, the users' attitudes to the conversational agents determined automatically during the conversations were compared to the results of a questionnaire taken after the conversations. The results provided by the system revealed similar tendencies to the questionnaire. Therefore we can say that the method is applicable as a means of evaluation for Japanese－speaking conversational agents.

Keywords：Affect－as－Information Theory, Conversational Agents, Evaluation Methods

1．Introduction

Technological development focused on enhancing and facilitating human lives has led to a need for intelligent environments meeting all human needs. Some examples are long-term projects, such as MIT's *House_n*[1], *MavHome*[2] or *Living Tomorrow*[3] Smart Home Projects. Explorations in the field of Ambient Intelligence (Ducatel et al., 2001) brought to light a new dimension of communication, where humans and machines become interlocutors, Human-Computer Interaction (HCI) (Dix et al., 2004). With this came a rush in development of intelligent conversational agents, beginning with freely talking chat-bots (Higuchi et al., 2008), through car navigation systems (Takahashi et al., 2003) to talking furniture (Hase et al., 2007). Their functional implementation into our lives has already become a current process. A need for an environment not only intelligent, but also humanized, is growing rapidly (Treur, 2007).

Along with this, researchers focused on agent development have found themselves with an urgent need to develop fast automatic evaluation methods for such agents. The usual methods used to evaluate conversational agents are based on subjective questionnaires in which user-testers express their opinions about the agent, their satisfaction during interaction with it, their will to continuing the conversation, the naturalness of the agent's utterance generation, etc. There have been some attempts to automatically evaluate spoken task-oriented dialogue systems, such as those by Litman, Walker and colleagues (Walker et al., 1997; Litman et al., 1998). However, these apply only to task-oriented spoken dialogue agents, and therefore are based on simple detection of keywords appropriate to the task performed by the English-speaking agent. A different approach was presented by Isomura and colleagues (Isomura et al., 2006), who made an attempt to evaluate a non-task-oriented Japanese-speaking dialogue agent using the Hidden Markov Model. However, their results were rather low

＊ Hokkaido University, Graduate School of Information Science and Technology, Language Media Laboratory

1 http://architecture.mit.edu/house_n
2 http://ailab.wsu.edu/mavhome/index.html
3 http://www.livtom.com/

(54%). Moreover, their method was able only to evaluate the naturalness of the agent's utterance, whereas in a usual subjective questionnaire there are many other dimensions other than naturalness in which the agent is evaluated. One could, for example, imagine a conversational agent that has very natural utterance generation, but through a lack of, e.g., rules of politeness, inappropriate proposition generation, or not keeping up the topic, would make the user irritated or even angry with the agent. Assuming that Isomura's method worked (54% of accuracy), such an agent would be evaluated in their method as being very good (very natural utterance generation); however, as the utterances eventually made the user dissatisfied, the overall evaluation would be rather negative.

To obtain a satisfying automatic evaluation method for conversational agents, we would thus need something that would act as a substitute for the subjective questionnaire. In questionnaire-type evaluation the users make decisions about how highly to mark the agent, and as such the process of questionnaire evaluation could be perceived from a typical decision-making perspective. The acts of decision-making and expressing opinions in humans strongly depend on features like emotional states (Loewenstein and Lerner, 2003; Rzepka and Araki, 2007). Therefore we assumed that it should be useful to analyze the attitudes of user-testers towards agents. Another problem with the questionnaire is that, as it is carried out after the conversations (sometimes an hour or more later, if the testing is time-consuming), the users' attitude may change from the time of the conversation. This change may be caused by the passing of time gradually obscuring the impression of the agent ; or, in the evaluation of two or more agents, the impression of the former may be altered by the performance (better or worse) of the latter; also, as is argued by Clore and colleagues (Clore et al., 2001; Clore and Storbek, 2006), changes in attitudes may be influenced by mood fluctuations caused by different factors, such as weather or, for example, news seen on television in the time between the actual experiment and filling in the questionnaire. All the above makes it most desirable to gather the attitudinal information from the users during the time of the conversation with the evaluated agent.

Ptaszynski and colleagues (Ptaszynski et al., 2008) were the first to propose such a method. During user conversations with two non-taskoriented Japanese-speaking conversational agents, they estimated the users' current attitudes and sentiments towards the agents. They based their idea on "Affect-as-Information" (Schwarz and Clore, 1983) reasoning about the emotions expressed in the users' utterances. However, one of the problems with this method was confusion of the valence polarity of emotive expressions in the last step of analysis performed by the affect recognition system used as a key tool in the method. To solve this problem, we applied the analysis of Contextual Valence Shifters (CVS) to the baseline system in order to enhance the specific emotion type determination. Moreover, the method was primarily tested only on a small number of evaluators. Therefore, we decided to test the method on a number of participants nearly three times larger than in the previous experiment.

The outline of this paper is as follows. In the second section we explain our approach to the task as a cooperation of sentiment and affect analysis. In Section 3 we provide all the necessary definitions of terms used frequently in this paper. Section 4 describes the main affect analysis system used in our method. Section 5 presents the description of the reasoning used to acquire important data from the affect analysis system output. In Section 6 we describe the settings of the experiment performed to verify the usability of the method, and in Section 7 we present and explain the results of this experiment. Section 8 presents discussion and interpretation of the results, and Section 9 contains concluding remarks. Finally, the paper closes with some ideas about how the method could be developed further and improved, which we plan to implement in the near future.

2. Our Approach–Attitude From Affect

2.1 Sentiment Analysis for Agent Evaluation

As mentioned above, in order to evaluate a conversational agent we need to obtain information about the user's attitudes toward the agent. The field focused on gathering such information is called Sentiment Analysis. It is a sub-field of Information Extraction that has only recently captured the interest of scientists (Turney, 2002). The general idea of sentiment

analysis is to gather and classify (into positive and negative) sentiments and attitudes about particular topics or entities. Sentiment Analysis is important for marketing research (Pang and Lee, 2008), monitoring of chatroom content for security reasons (Abbasi and Chen, 2007), and customer feedback on particular products (Turney, 2002). As we can consider that conversational agents are ultimately products as well, it would be desirable to acquire objective information about the agents' performance before putting them on the market, as failure may cause a substantial loss of funds and human effort. Tests, where people are hired to verify the performance of market-destined agents, are burdened with heavy use of effort and funds. Moreover, paying user-testers high sums of money for the evaluation undermines the objectivity of such a test. Although there is no other way of performing the test than making a human talk to an agent, in our assumption there is a better way to gather more objective information for the evaluation than a typical questionnaire performed after the test phase. Namely, information about the tester's sentiment towards the product (agent) could be gathered during the test phase (conversation with the agent). This should provide the objective information. However, in the usual sentiment analysis methods, the attitudinal information is extracted from the text with regards to a particular object (product). This means that such methods are applicable only if the user explicitly express their attitude towards the product. Unfortunately, in a free, non-task-oriented conversation, users usually do not express their attitudes directly towards their machine interlocutors. However, we worked using the assumption that the users' attitude should be revealed in how they respond to the agents' utterances. Therefore, analyzing the emotional level of users' utterances and applying some kind of function transforming this data into attitudinal information should provide us with information about what users think about the agents. This, if mapped efficiently on a set of questions from a usual subjective questionnaire, would in effect provide a substitute for the questionnaire.

We decided to gather information about users' emotional states during conversations using one of the techniques for Affect Analysis, and transform the data obtained this way using reasoning based on the "Affect-as-Information" Theory to determine the user's attitude towards the interlocutor, the agent.

2.2 Affect Analysis for Attitude Estimation

Affect Analysis is also a relatively new sub-field of Information Extraction, and focuses on classifying users' expressions of emotions. However, in contrast to Sentiment Analysis, where the goal is to determine the user's general attitude (positive or negative) to a specific object (movie review, or a product), this field takes as an object the user himself, and its goal is to estimate human emotional states in a more detailed manner. While attitude could be either positive or negative, the expression of emotion could represent a wide scope of emotional states, from fear, anger, or excitement to joy, pleasure, or relief. In the most popular methods, emotions are determined from facial expressions (Hager et al., 2002), voice (Kang et al., 2000) or biometric data (Teixeira et al., 2008). However, as emotions are context dependent (Mandel, 2003) and their semantic and pragmatic diversity is best conveyed in language (Solomon, 1993), most of the semantic content of expressing emotions is ignored in such research. Therefore, we decided to use a method to analyze the affect of textual representation of an utterance. There is some research on affect analysis, also for the Japanese language (Tsuchiya et al., 2007; Ptaszynski et al., 2008). However, there have been only a few approaches to apply affect analysis to gather information about sentiments and attitudes (Grefenstette et al., 2004), and no significant work has been done on applying such an approach to the evaluation of conversational agents using Japanese. This paper presents the first attempt of this kind.

3. Definitions

3.1 Definition and Classification of Emotions

Nakamura (Nakamura, 1993) defines emotions as every temporary state of mind, feeling or affective state evoked by experiencing different sensations. This definition is complemented by Solomon, who argues that people are not passive participants in their emotions, but rather that emotions are strategies by which people engage with the world (Solomon, 1993). With regard to language, the above is further complemented by Beijer's definition of emotive utterances, which he describes as every utterance in which the speaker is emotionally involved, and this involvement,

expressed linguistically, is informative for the listener (Beijer, 2002).

Nakamura proposed a 10-type classification of emotions : *ki / yorokobi*[4] (joy, delight), *do / ikari*(anger), *ai / aware* (sorrow, sadness, gloom), *fu / kowagari* (fear), *chi / haji* (shame, shyness, bashfulness), *kou / suki* (liking, fondness), *en / iya* (dislike, detestation), *kou / takaburi* (excitement), *an / yasuragi* (relief) and *kyou / odoroki* (surprise, amazement). We used this classification instead of the common practice of proposing our own, as Nakamura's several decades-long research on emotive expressions makes his classification the most appropriate for the Japanese language.

3.1.1 Clarifying the Nomenclature

In this subsection we briefly clarify the differences between some of the emotion-related terms used in this paper.

Emotion The classic definition of **emotion** says that it is a mental and physiological state caused by subjective experiences. However, in modern psychology and cognitive science (Solomon, 1993) it is perceived more as a process in time including various specifically defined phenomena, such as affective states, sentiments, moods, or changes in attitudes (see also above for our working definition of emotion).

Feeling is defined in psychology as a conscious subjective experience of any physical sensation (VandenBos, 2006). In common sense understanding it is used not only in terms of emotions, but includes also other sensations, such as "warm", "cold" or "soft" -also subjective and evaluative, but not directly emotional.

Affect is often referred to as the experience of feeling (Huitt, 2003) and represents an organism's reaction to stimuli. **Affective state** is the state caused by the experience of feeling (affect) and includes a process during which the organism interacts with and responds to the stimuli. The linguistic part of this phenomenon, on which we focus in particular, includes expressing one's emotions in a way informative to the environment (other interlocutors, such as people or an agent).

Mood is usually distinguished from affect on the basis of time and intentionality. It is said to be a relatively long-lasting emotional state not caused by any easily determinable stimuli. It is known that moods can be caused by changes of weather or diet. It was also discovered that moods influence people's tendencies in decision-making (Schwarz and Clore, 1983).

Sentiment is defined as a person's conscious opinion, or attitude tendency towards an object. In the context of Sentiment Analysis it refers to attitudes (positive or negative sentiments) or opinions (specific).

Attitude in psychology refers to a person's perspective toward a specified object, in particular one's degree of liking or disliking of the object (Breckler and Wiggins, 1992).

3.2 Two-dimensional Model of Affect

The idea of a two-dimensional model of affect was first proposed by Schlosberg (Schlosberg, 1952) and was developed further by Russell (Russell, 1980). Its main assumption is that all emotions can be described in a space of two dimensions: the emotion's valence (positive/negative) and activation (activated/deactivated). An example of positive-activated emotion would be "excitement"; a positive-deactivated emotion is, for example, "relief"; negative-activated and negative-deactivated emotions would be "anger" and "gloom" respectively. In this way, four areas of emotions are distinguished: activated-positive, activated-negative, deactivated-positive and deactivated-negative (see Figure 1).

Nakamura's emotion types were mapped on this two-dimensional model of affect, and their affiliation to one of the spaces was determined. The emotion types for which affiliation is not obvious (e.g. surprise can be both positive as well as negative; dislike can be either activated or deactivated, etc.) were mapped on all of

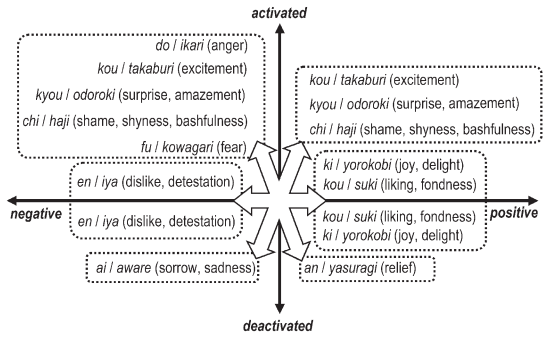

Fig.1 Grouping Nakamura's classification of emotions on Russell's space.

4 In this paper we use italics for expressions in Japanese.

the groups they could belong to. However, no emotion type was mapped on more than two adjacent fields. This grouping is then used in our system for two purposes. First, in the CVS analysis procedure, the grouping is used to specify which emotion corresponds to the one negated by a CVS phrase. Second, it is used to estimate the attitude towards an agent by determining the valence polarity of emotions conveyed during a conversation.

3.3 Affect-as-Information Theory

The theory of Affect-as-Information was introduced in 1983 by Schwarz and Clore (Schwarz and Clore, 1983) and is widely studied in the field of psychology and social psychology. Schwarz and Clore claimed that people use affect in the same way as any other criterion, by applying the informational value of their affective reactions to form their judgments, attitudes and opinions.

Schwarz, Clore and colleagues studied this phenomenon thoroughly in numerous experiments (Schwarz and Clore, 1983; Clore et al., 2001; Clore and Storbek, 2006). They reached the conclusion that people's choices and evaluations, and therefore attitudes, change according to the changes in their current moods. This change could be caused naturally (e.g. weather), or induced by various factors. For example, watching a sad movie induces sad moods in a person, which could be further used as a factor to cancel a party with friends. As another example, talking to someone one hates may spoil the whole day. Similarly, talking to someone interesting and friendly could induce positive mood, and the overall estimation of one's relationship with this person could be even better.

Using the same reasoning, we assumed Schwarz and Clore's findings to be useful in transforming the results of affect analysis of user utterances carried out during conversation with an agent into information about the users' attitudes towards the evaluated agents. The subsequent filling in of a subjective questionnaire about the interlocutor after the conversation can be perceived as a typical decision-making process (people make decisions about how to evaluate the agent). Therefore, if the approach is correct, the automatic estimation of users' attitudes through the conversation should indicate similar tendencies to the results acquired through the questionnaire.

Proving this to be true would be a step towards the practical realization of the idea of affective human factors design (Jiao et al., 2007), where the information about a product (agent) is derived from information about dynamic changes of the user's affective states during usage. If proved, this would provide strong evidence that in the process of product design, affective factors are not only as important as usability (Jiao et al., 2007), but that affect itself provides valuable information about usability, and can thus be a source of information for continuous improvement of the product.

3.4 Contextual Valence Shifters

The idea of Contextual Valence Shifters (CVS)' application in Sentiment Analysis was first proposed by Polanyi and Zaenen (Polanyi and Zaenen, 2004). They distinguish two kinds of CVS: negations and intensifiers. The group of negations contains words and phrases like "not", "never", and "not quite", which change the valence polarity of the semantic orientation of an evaluative word they are attached to. The group of intensifiers contains words like "very", "very much", and "deeply", which intensify the semantic orientation of an evaluative word. So far the idea of CVS analysis was successfully applied to the field of Sentiment Analysis of texts in English (Kennedy and Inpken, 2005). A few attempts by Japanese researchers (Miyoshi and Nakagami, 2007) show that it is also applicable for the Japanese language.

Examples of CVS negations in the Japanese language are grammatical structures such as *amari -nai* (not very-), *-to wa ienai* (cannot say it is-), *mattaku -nai* (not at all-), or *sukoshi mo -nai* (not even a bit-). Intensifiers are represented by such grammatical structures as *totemo-* (very much-), *sugoku-* (-a lot), or *kiwamete-* (extremely).

4. ML — Ask — Affect Analysis System

The affect analysis system employed in the automatic evaluation method described in this paper is ML-Ask developed by Ptaszynski et al. (Ptaszynski et al., 2008; Ptaszynski et al., 2009b). To realize the method, one can use any reliable affect analysis system available in the field. However, as mentioned in section 2, the information about users' affective states needs to

be extracted and analyzed in real time. Therefore, we used Ptaszynski's system as it is fast (analysis of one utterance takes less than 0.15 seconds) and reliable (different evaluations confirmed the system's reliability in laboratory conditions as well as in the field; for details see: (Ptaszynski et al., 2008; Ptaszynski et al., 2009b; Ptaszynski et al., 2009c)).

ML-Ask (eMotive eLements / Emotive Expressions Analysis System) was developed for analyzing the emotive contents of utterances. The system uses a two-step procedure: 1) Analyzing the general emotiveness of an utterance by detecting emotive elements, or emotemes, expressed by the speaker and classifying the utterance as emotive or non-emotive; 2) Recognizing the particular emotion types by extracting expressions of particular emotions from the utterance. This analysis is based on Ptaszynski's (Ptaszynski, 2006) idea of two-part classification of realizations of emotions in language into:

1) *Emotive elements or Emotemes.* Elements conveyed in an utterance indicating that the speaker was emotionally involved in the utterance, but not detailing the specific emotions. The same emotive element can express different emotions depending on context. This group is linguistically realized by subgroups such as interjections, exclamations, mimetic expressions, or vulgar language. Examples are: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (a vulgarization of a verb);

2) *Emotive expressions.* Parts of speech used to describe emotional states. However, they function as expressions of the speaker's emotions only in utterances where the speaker is emotionally involved. In non-emotive sentences they fulfill the function of simple descriptive expressions. The group is realized by various parts of speech, like nouns, verbs, adjectives, etc. Examples are: *aijou* (love), *kanashimu* (feel sad), *ureshii* (happy), respectively.

The emotive element database was hand-selected using data from different research (Oshima-Takane et al., 1995−1998; Tsuchiya, 1999; Baba, 2003; Sjöbergh, 2006) and divided into interjections, mimetic expressions, endearments, vulgarities, and representations of nonverbal emotive elements, such as exclamation marks or ellipses. The emoteme database collected and divided in this way contains 907 elements in total. A simple algorithm detecting emoticons was also added, as they are symbols commonly used in everyday text-based communication tools. The database of emotive expressions is based on Nakamura's collection (Nakamura, 1993) and contains 2100 emotive expressions, each classified into the emotion type they express.

4.1 Affect Analysis Procedure

On textual input provided by the user, two features are computed in order: the emotiveness of an utterance and the specific type of emotion.

To determine the first feature, the system searches for emotive elements in the utterance to determine whether it is emotive or non-emotive. In order to do this, the system uses MeCab for morphological analysis and separates every part of speech (Kudo, 2001). MeCab recognizes some parts of speech we define as emotemes, such as interjections, exclamations or sentence-final particles, like *-zo*, *-yo*,or *-ne*. If these appear, they are extracted from the utterance as emotemes. Next, the system searches and extracts every emoteme based on the system's emoteme databases (907 items in general). Finally, the simple emoticon detector informs about the presence of emoticons in the utterance. This is performed by detecting the appearance of at least three symbols in a row, used usually in emoticons. A set of 362 of those symbols was selected as being the most frequent symbols appearing in emoticons analyzed by Ptaszynski (Ptaszynski, 2006). All of the extracted elements mentioned above (exclamations from MeCab, emotemes and emoticons) indicate the emotional level of the utterance.

Table 1 An example of analysis performed by ML-Ask (system output). Emotemes - underlined ; emotive expressions - bold type font ;

Utterance	*Iyaa, kyou wa nante* ***kimochi ii*** *hi nanda!* ^o^ (Oh, today is such a nice day! ^o^)
Emotive elements	exclamative sentence structure: *nante–nanda* interjection: *iyaa* emotive mark: *!* emoticon: ^o^
Emotive expression	***kimochi ii*** (nice [feeling])

Secondly, in utterances classified as emotive, the system uses a database of emotive expressions to search for all expressions describing emotional states. This determines the specific emotion type (or types) conveyed in the utterance. An example of analysis performed by ML-Ask (system output) is shown in Table 1.

4.2 CVS Procedure in ML-Ask

One of the problems in the procedure described above was confusion of the valence polarity of emotive expressions. The cause of this problem was extracting from the utterance only the emotive expression keywords without its grammatical context. One utterance showing such a case is presented in Table 2. In this sentence the emotive expression is the verb *akirameru* (to give up [verb]) and a CVS phrase, *-cha ikenai* (Don't-[particle+verb]) is present, suggesting that the speaker is in fact negating and forbidding the emotion expressed literally. To solve the problem we applied the analysis of Contextual Valence Shifters to change the valence polarity of emotive expressions in utterances containing CVS structures.

However, using only the CVS analysis we would be able to find out about the appropriate valence of emotions conveyed in the utterance, but we would not know the exact emotion type. Therefore, to specify the emotion types in such utterances we applied the idea of the two-dimensional model of affect.

Table 2 An example of failure in emotion deter - mination in ML-Ask and improvement by CVS. Emotemes - underlined ; emotive expressions - bold type font ; right arrow above - CVS structures.

	Human annotation	ML-Ask baseline	ML-Ask + CVS
Akirame $\overrightarrow{cha ikenai}$ *yo!* (Don't ya give up!)	[joy],	[dislike]	[joy], [liking]
Sonna ni ***omoshiroku*** $\overrightarrow{monakatta}$ *yo...* (Oh, it wasn't that interesting...)	[dislike, boredom]	[joy]	[dislike]

4.2.1 Applying Two-dimensional Model of Affect to CVS Procedure

The need to change the valences in emotion estimation research is a common problem. However, it is not uncommon for researchers to use valence changing patterns constructed by themselves without any scientific grounds. For example Tsuchiya and colleagues (Tsuchiya et al., 2007) used their own list of contrasting emotions. However, they do not consider that, as is argued by Solomon (Solomon, 1993), the fact that two emotions are in contrast is not a matter of clear division, but is more complex and context dependent. We assumed this complexity could be specified with the help of the two-dimensional model of affect.

4.2.2 Description of CVS Procedure

Analysis of Contextual Valence Shifters is a supplementary procedure for ML-Ask and works as follows. When a CVS structure is discovered, ML-Ask changes the valence polarity of the emotion conveyed in the sentence. Every emotion is placed in a suitable space according to Russell's model. The appropriate emotion is determined as belonging to the emotion space with both valence polarity and activation parameters opposite to those of the primary emotion (note arrows in Figure 1). If an emotion was located in only one quarter, e.g. positive-activated, the contrasting emotions would be determined as negative-deactivated. A change in the output is shown in Table 2. In the first example, originally ML-Ask selected [dislike]. This emotion is located in both quarters of the negative valence space. Therefore, after valence shifting, ML-Ask determines the new emotion types as positive and belonging to both of the positive quarters. The new proposed emotion types are: [joy] and [liking] belonging to both positive-activated and positive-deactivated quarters. The second example presents the opposite situation. The procedure, as described above, was shown to improve affect analysis in Japanese by Ptaszynski and colleagues (Ptaszynski et al., 2009a). The system flow chart including CVS procedure is shown in the upper part of Figure 2.

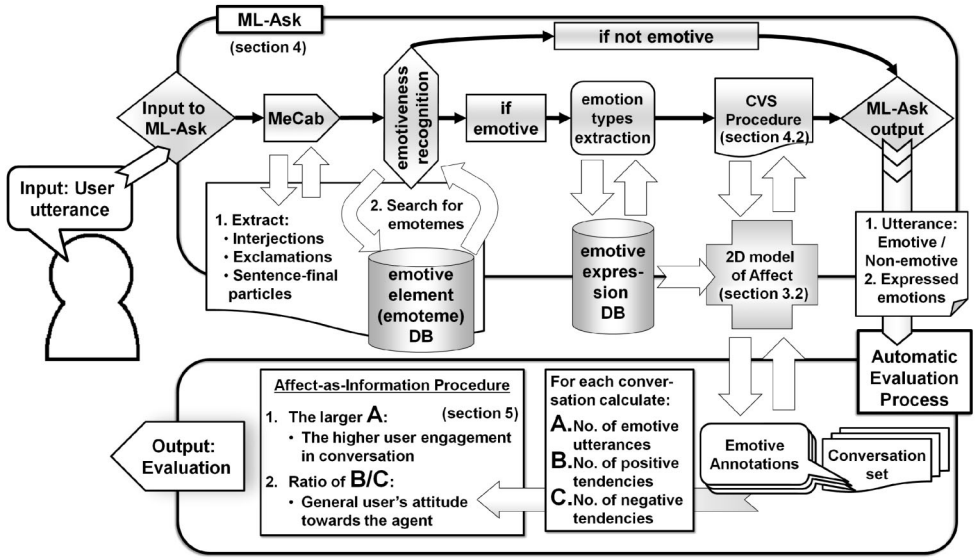

Fig.2 Flow chart of the automatic evaluation procedure including : affect analysis system ML-Ask with CVS procedure (upper part); further processing of information obtained by ML-Ask and the decision making process for the final evaluation.

5. Information Derived from ML-Ask Output

ML-Ask is used to analyze utterances of a user talking to a conversational agent, during the conversation. The results of analysis of each utterance provide information on how many user utterances were emotive. Furthermore, the emotions extracted from the user's emotive utterances form a vector on which the emotional states of the user changed during the conversation. This is then processed as follows.

Firstly, if many[5] of the user's utterances were determined as emotive, we assume the user was emotionally involved in the conversation. Emotional involvement in a conversation suggests a tendency towards easier familiarization with the interlocutor (Yu et al., 2004). Therefore we can assume that during a user's conversation with an agent, the machine interlocutor is considered to be more human-like the more emotionally emphasized the user's utterances are. However, this does not yet mean a positive familiarization. The conversation could become emotional also when the interlocutors quarrel. This could happen in a case where the agent makes the user angry. However, if the user agrees to participate in a quarrel with an agent, this could also mean that the user finds the agent's linguistic capabilities to be comparable to himself. Therefore, the information obtained about the general emotiveness of the conversation could be interpreted as signifying how much the user finds the agent worth talking to, including familiarity and the user's opinion about the agent's linguistic skills.

Secondly, analysis of specified emotion types conveyed by the user in the whole conversation provides information on the user's particular emotions during the conversation. If the emotions according to the Russell's model were positive or changing from negative to positive while talking, the general attitude towards the agent is considered to be positive. If the emotions were negative or changing from positive to negative, the attitude is classified as negative. The general attitude towards an agent is calculated as the ratio of conversations with positive tendency to the conversations with negative tendency. The flow of the procedure is presented in Figure 2.

Both types of information acquired (general engagement of the user in conversation and attitude) provide an overview of the user's sentiment about the agent,

5 In this paper we do not specify the value of "many". We compare the results for two different conversational agents to verify for which the tendency was higher. However, we assume it is possible to set a threshold in the results for evaluation of only one agent. Such a threshold could be obtained statistically, after performing several evaluations of different agents. It could also be set arbitrarily, as, for example, in (Landis and Koch, 1977).

and it is desirable for both types of information to harmonize rather than show dissonance. Such analysis, if accurate, realizes the first step of affective human factors design, which is to understand the user's affective needs (Jiao et al., 2007).

6. Evaluation experiment

To test our method, we performed an evaluation experiment of two non-task-oriented conversational agents. The first agent is a simple conversational agent which generates responses by 1) using Web-mining to gather associations to the content of user utterance; 2) making propositions by inputting the associations to the prepared templates; and 3) adding modality to the basic propositions to make the utterance more natural. The second agent, based on the first one, generates a humorous response to user utterance every third turn. The humorous response is a pun created by using user input as a seed to gather pun candidates from the Web and inputting the most frequent ones into pun templates (for more detailed description of the agents see below and references). The choice of the agent was deliberate. They differed only in one respect -the humorous responses in the latter one. The assumption was that, as humor is an important factor in socialization (Yup and Martin, 2006), the joking agent should be evaluated higher by the users and this difference should be easily seen. If the automatic evaluation method then displayed the same tendencies, they should also be easily recognizable.

There were 13 participants in the experiment, 11 males and 2 females. All of them were university undergraduate students. The users were asked to perform a 10-turn conversation with both agents. No topic restrictions were made, so that the conversation could be as free and human-like as possible. The agents were first evaluated during the conversation using the proposed automatic evaluation method and the results were stored for further comparison with a subjective questionnaire. After the conversations, the users were asked to complete a questionnaire concerning their attitudes to the agents and their performance. The results of the automatic evaluation were compared to the results of a subjective questionnaire filled in by the users in order to evaluate the two agents. Using these sets of results, we were looking for similarities between sentiment classification and the questionnaire.

6.1 Two Conversational Agents -Short Description

Modalin

Modalin is a non-task-oriented text-based conversational agent for Japanese. It automatically extracts from the Web sets of words related to a conversation topic set freely by a user in his utterance. The association words retrieved from the Web (with accuracy of over 80%) are then sorted by their co-occurrence on the Web, and the most frequent ones are selected to be used further in output generation. In response generation, the extracted associations are put into one of the pre-prepared response templates. The choice of the template is random, but the agent keeps in its memory the last choice in order not to generate two similar sentence patterns in a row. Finally, the agent adds a modality pattern to the sentence and verifies its semantic reliability. The modality is added from a set of over 800 patterns extracted from a chat-room logs and evaluated. The naturalness of the final form of the response is then verified on the Web with a hit-rate threshold set arbitrary for 100 hits. The agent was developed by Higuchi and colleagues. For further details see (Higuchi et al., 2008).

Pundalin

Pundalin is a non-task-oriented conversational agent for Japanese, created on the base of Modalin combined with Dybala's Pun generating system PUNDA (Dybala et al., 2008b). The PUNDA pun generator was developed by Dybala and colleagues as a part of PUNDA research project, aiming to create a Japanese pun generating engine. The system work as follows. From the user's utterance, a base word is extracted and transformed using Japanese phonetic pun generation patterns, to create a phonetic candidate list. The candidate with the highest hit-rate in the Japanese search engine Goo[6] is chosen as the most common word that sounds similar to the base word. Next, the base word and the candidate are integrated into a sentence. The integration is done in two steps, one for each part of the sentence including the base word and the pun candidate, respectively. Firstly, the base phrase is put into one of several pre-prepared templates making up the first half of the sentence. The second half of the sentence is extracted from KWIC

6 http://search.goo.ne.jp/

on WEB -on-line Keyword-in-context sentences database (Yohsihira et al., 2004) as the shortest latter half of an emotive sentence including the candidate. Every third turn of the conversation, Modalin's output was replaced by a joke-including sentence, generated by the pun generator. Pundalin therefore is a humor-equipped conversational agent using puns to enhance communication with the user. Pundalin was developed by Dybala and colleagues as a conversational agent for use in experiments on the influence of humor on human-agent interaction (Dybala et al., 2008a).

6.2 Questionnaire -User's Evaluation

The questions we asked users after the conversations with both agents were: A) Do you want to continue the dialogue?; B) Was the agent's conversation grammatically natural?; C) Was the agent's conversation semantically natural?; D) Was the agent's vocabulary rich?; E) Did you get an impression that the agent possesses any knowledge?; F) Did you get an impression that the agent was human-like?; G) Do you think the agent tried to make the dialogue more funny and interesting? and H) Did you find the agent's talk interesting and funny?. The answers for the questions were given in 5-point scale (1 -the lowest score ; 5 the highest score) with some explanations added. Each user filled two such questionnaires, one for each agent. The final, summarizing question was "Which agent do you think was better?"

6.3 Representation of Questionnaire in Sentiment Analysis

We made the following assumptions about how the questions we asked users directly were represented by the results provided by the analysis. We assumed that the questions from A) to H) generally represent several kinds of information, such as: how highly did the users evaluate agents' talking abilities (questions A-D); how much the users were able to familiarize with the agents (questions E-F); and how much they were emotionally involved in the conversation (questions G-H). According to Dybala (Dybala et al., 2009), in the evaluation of conversational agents there are two features that have to be evaluated. The first represents the agent's linguistic capabilities, and the second represents all features other than linguistic, such as subjective impression or ease of familiarization. In our assumption, the first set of questions (A-D) inquire about the linguistic features and the latter two sets of questions (E-F and G-H) represent the non-linguistic features. Further, the general summarizing question represented the users' general attitude towards the agents, and therefore represents the second type of information obtained from the automatic analysis (for details see Section 5).

7. Results

The results of the evaluation are shown below. First, the results of the questionnaire are shown, then the results of the automatic evaluation method are summarized and compared to the users' direct opinions.

7.1 User Evaluation

Regarding the detailed questions, higher scores were given by the users to Pundalin (see Figures 4 Table 3 with its graphical representation in Figure 5). Although the differences between Modalin and Pundalin were not that obvious in all categories, overall results for both agents clearly showed that the performance of Pundalin was estimated as being more human-like and easier to familiarize with.

The questions about agents' conversational abilities (questions B-D) revealed that the humor-equipped agent was rated higher, although the differences were not as great as in other questions. The reason for this is that Pundalin was based on Modalin and, with the exception of the humorous responses, all other were made in the same way as in Modalin. The questions inquiring how easily the users could familiarize with the agents (A and E-F) showed that Pundalin scored higher here as well. The most notable differences were seen in the questions investigating how much the users were emotionally involved in the conversation (questions A and G-H), where the joking agent was also evaluated higher. The results were summarized for all questions (with approximated values for users) in Table 3. All of the results were statistically significant at 5% level, except questions A and B, which were significant at 6% and 7% level respectively. The overall compared results of Modalin and Pundalin were extremely statistically significant, with P value = .0002. We also summarized the results for all users (with approximated values for questions), which also

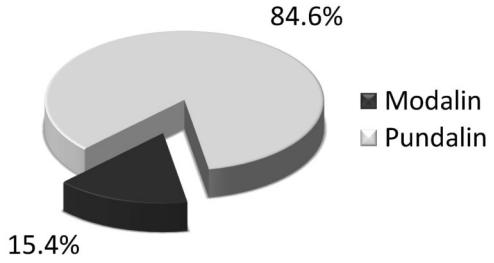

Fig.3 Users' evaluation-results for the question "Which agent do you think was better?".

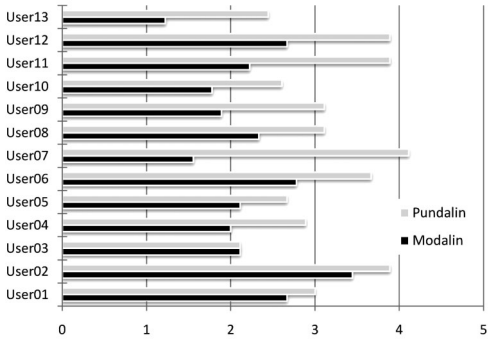

Fig.4 Users' evaluation for Modalin and Pundalin, representing the approximated results of all detailed questions per user. Answers given in a 5-point scale.

Table 3 Users' overall evaluation of Modalin and Pundalin for each detailed question. Answers given in a 5-point scale.

Questions	A	B	C	D	E	F	G	H
Modalin	2.62	2.15	1.85	2.08	2.15	2.38	1.92	2.46
Pundalin	3.38	2.92	2.69	3.00	2.85	3.31	4.15	4.08
P-values	.0544	.0646	.0205	.0160	.0323	.0045	.0005	.0035

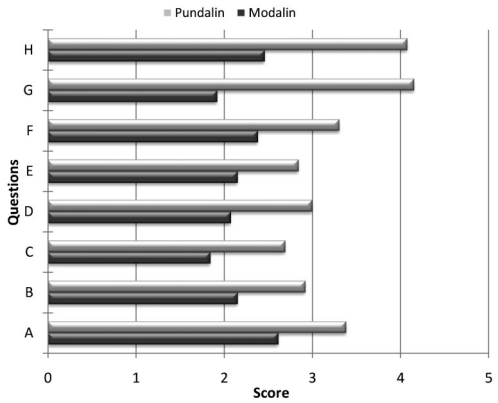

Fig.5 Graphical representation of Table 3. Results for each detailed question per agent. Answers given in a 5-point scale.

showed clearly that the users generally evaluated the joking agent higher (see Figure 4). These results were also extremely statistically significant, with P value = .0002.

This corresponds to the results of the final question, in which the users were asked which agent was better in general. This question investigated the general attitude of the users towards each agent after the experiment. 11 out of 13 users (84.6%) evaluated Pundalin (humorequipped agent) as better than Modalin (see Figure 3), which means the attitude was more positive towards the former agent.

After gathering the results of the questionnaire, we compared them to the automatic evaluation method. We assumed that if the tendencies were similar and the results were statistically significant, the method is applicable as an automatic evaluation method for non-task-oriented conversational agents.

7.2 Results of Sentiment Analysis

Evaluation based on sentiment analysis of the users' utterances showed tendencies similar to the questionnaire. The users were more emotionally involved in the conversations with Pundalin, which corresponds to the direct opinions about the agent -that it was more human-like, its utterances were more correct semantically, grammatically, etc. (see Figure 7) and therefore the agent was easier to familiarize with. The results summarized for all users were very statistically significant (P value = .0053).

The analysis of specified emotion types conveyed by the users in conversations provided information

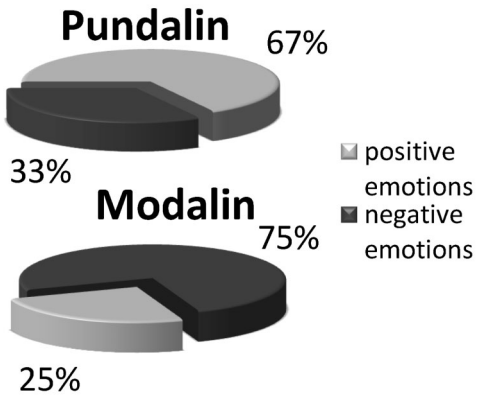

Fig.6 The total ratio of all emotions positive to all negative conveyed in the utterances of users with Modalin and Pundalin.

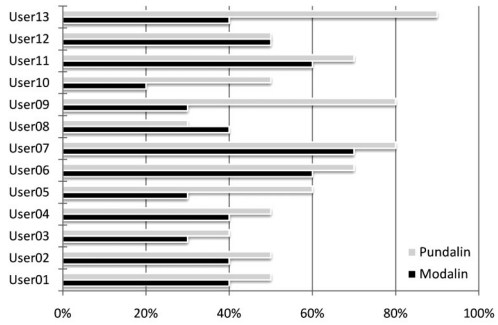

Fig.7 Average appearance of emotively engaged utterances for all 13 users in conversations with both agents ("90%" means that in 10-turn conversation there were 9 emotive utterances).

clearly revealing the users' attitudes towards each agent. The users' general attitudes to Pundalin were mostly positive (67%), whereas to Modalin the attitudes of the users were mostly negative (75%). For details see Figure 6.

The results above indicate that the general attitude of a user towards an agent was better for Pundalin than for Modalin, which corresponds to the results of the questionnaire.

7.3 Correlations Between Automatic Evaluation and Questionnaire

In order to check which questions were correlated best with the results of automatic evaluation, we calculated the correlation coefficient. As the base for calculations we used Spearman's rank correlation coefficient (Spearman's ρ [rho]). The usual Pearson's correlation coefficient represents a linear dependence between data, which is not the issue in subjective evaluation. Therefore Spearman's rank correlation coefficient used to calculate any monotonic dependence is more appropriate for our task. The results are presented in Table 4.

In this research we aim to propose a method to substitute the usual subjective evaluation questionnaire, and thus the most important particular question sets for us were those representing non-linguistic features (especially G and H). The correlation test revealed accordingly that the strongest correlation was between sentiment analysis and questions G (Did the agent try to be interesting?) and H (Was the agent interesting?). Therefore, we can say that the automatic evaluation method is applicable in subjective evaluation of non-linguistic features, especially those related to entertaining the user.

Table 4 The results of Spearman's rank correlation test between sentiment analysis and each question.

Question	A	B	C	D	E	F	G	H
ρ	**.333**	**.350**	.202	.164	**.480**	.035	**.559**	**.597**

The test revealed also other correlations. A medium-strong correlation was found in the results of question E (Did the agent posses any knowledge?). This can be interpreted to signify that people usually become more involved in conversation with intelligent interlocutors. Medium correlation was also found with questions A (Continuing the dialogue) and B (Grammatical naturalness). The first can be interpreted as a natural consequence of the results for question E stronger involvement in the conversation with an intelligent partner logically makes one more obligated to continue the dialogue. The cause of respectively high correlation of B is not visible at first glance, but when set together with questions C (Semantic naturalness) and D (Vocabulary richness) becomes more understandable. The correlation of these questions with the automatic evaluation declines along with an increase in possibilities of interpretation. This is presumably also the reason for question F (Human-likeness) to be the least correlated, since, as noted also by Dybala and colleagues (Dybala et al., 2010), the concept of human-likeness in machines is still vague and undefined.

It is also possible that changing the formulation of the questions and improving the method itself will enhance the correlation as well. Moreover, there already exist automatic evaluation methods for linguistic abilities of conversational agents (Isomura et al., 2006), although their accuracy is not high. However, combining them with our method might show improvement of the overall evaluation.

8. Discussion

In the primary evaluation experiment of this method, performed on two conversational agents, Ptaszynski and colleagues (Ptaszynski et al., 2008) showed, using five user-testers in their experiment, that there were similar tendencies in the results acquired by the method and the results of the questionnaire.

The results presented here, although the number of evaluators was nearly three times larger (13 participants), show that the tendencies remained the same. Users showed higher emotive engagement and positive attitudes in conversations with the agent which used jokes. This proves that the method is applicable as a means of evaluation for conversational agents.

The differences between results of the questionnaire and the method were not in a one-to-one ratio, however, it should be remembered that both evaluations, although aiming to provide answers to similar questions, were based on different assumptions. In the questionnaire the users are aware of the points they deliberately assign, whereas in the automatic evaluation method the users do not know that what they say will be used in evaluating the agent. Compared to traditional subjective questionnaires, this makes the proposed method less invasive and therefore provides objective information on the users' sentiments about the machine interlocutor.

The automatic evaluation correlated strongest with the questions about non-linguistic features. As there has not previously been a method for automatic evaluation of such features, this is probably one of the most significant achievements of this method. The questions about linguistic abilities also correlated, although in a weaker manner. However, we can predict that improving the method, either by improving the intermediary procedures or by combining it with other automatic evaluation methods, will improve the overall evaluation. Moreover, the representation of sentiment analysis results in the questions was set arbitrarily and it is possible that there could be a set of questions which represents the information obtained by automatic evaluation in a more straightforward manner. However, for the experiment presented in this paper, the attention should rather be focused on the similarities in tendencies that appeared in general comparison of the two agents and on the fact that all compared results were statistically significant.

Approximate time of processing one utterance is below 0.15 s, which makes the method applicable in providing actual information on changes in the users' attitudes towards the machine in real time. This does not only provide fast and up-to-date information on users' sentiments, but also, appropriately utilized, can provide hints for the agent about potential undesirable changes in the users' attitudes and the need for appropriate counteractions, during everyday use.

9. Conclusions

In this paper we presented an automatic method of evaluation for conversational agents. The method is based on analyzing affective states conveyed by a user in a conversation with an agent. Borrowing the notion of affect-as-information (Schwarz and Clore, 1983), the results of affect analysis performed by a system created by Ptaszynski and colleagues (Ptaszynski et al., 2008; Ptaszynski et al., 2009b) provide us with information about the user's emotional involvement in a conversation, closing of psychological distance, and ease of familiarization with the machine. This corresponds to direct questioning of the user about the agent's performance. Next, analysis of specified emotion types conveyed by the user in the whole conversation and their classification by applying the two-dimensional model of emotions (Russell, 1980) provides us with information on the polarity of the users' attitudes towards the machine interlocutor during the conversation.

By applying the proposed method in evaluation of conversational agents, the evaluating information is acquired in the process of testers' conversing with an agent. Therefore as means of evaluation, the method saves time, effort and funds spent each time on preparing and performing laborious questionnaires. It is desirable for the proposed method to be accepted widely in the field as a full equivalent or at least a strong supportive means to objectivize the results of traditional questionnaires.

10. Future Work and Perspectives

Our method, although proven to be effective, still has still some deficiencies which we aim to rectify in the near future. The imperfections of the subsystems used in the method influence its accuracy. The slight deficiency in the emotion types extraction procedure in ML-Ask limit the information about affective states conveyed by users in conversation. However, we can predict that applying the two-dimensional model of emotions into assigning emotional affiliations of emotive elements will disambiguate the emotional affiliations of emotive elements, thus improving the

performance of ML-Ask. Some ideas about ways to improve the system were already proposed by Ptaszynski and colleagues (Ptaszynski et al., 2009d). We plan to implement them in the near future.

The method should be also tested on other agents than the two presented here. Dybala and colleagues, after adding some improvements mentioned above, have already used our method to evaluate two different conversational agents (Dybala et al., 2010). However, the differences between their agents were similar to the two agents compared in this paper -one was a simple conversational agent (HMM based) and the second one used jokes, although the appropriate timing for joke generation was not set arbitrarily every third turn, like here, but was based on analysis of the emotional states of the users. Therefore, it is desirable to verify the usability of our automatic evaluation method also on conversational agents which differ in features other than the generation of humorous responses.

The method is designed for the Japanese language, although constructing a different language version of ML-Ask would be possible after gathering adequate databases for other languages, especially ones with similar morphology, like Korean. The notion of affect-as-information, although with a firm scientific basis in psychology and social psychology (Clore et al., 2001; Clore and Storbek, 2006), is not a common notion in the fields we referred to in this paper, Agent Development, Evaluation Methods, Affect Analysis, or Artificial Intelligence in general. The mapping of questions on the results of affect analysis, although supported with strong theory, is still rather commonsensical and intuitive. Therefore, we will aim to make the mapping more precise in future by looking for the questions that correlate strongest with the automatic evaluation. However, in this experiment we tried to prove that affective states do influence judgments and attitudes towards agents and, properly analyzed, reveal similar tendencies to usual evaluation questionnaires, providing valuable and important information in evaluation - a significant part of the product design process. Coordinating the appropriate items for automatic evaluation with the questions asked directly to the user is based on psychological reasoning, and therefore reaches deeper and beyond the simple numbers usually put in terms of familiar notions of accuracy or precision and recall. However, the rapid development in all fields of science, as well as in commercial areas, compels researchers from different scientific fields to join efforts, which - as we have shown in this paper - can be successful.

Acknowledgment

This research was 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.

Baba, Junko. "Pragmatic function of Japanese mimetics in the spoken discourse of varying emotive intensity levels." *Journal of Pragmatics*, Elsevier. 2003.

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

Breckler, S. J., and Wiggins, E. C."On defining attitude and attitude theory: Once more with feeling". In A. R. Pratkanis, S. J. Breckler, and A. C. Greenwald (Eds.). *Attitude structure and function*. Hills-dale, NJ: Erlbaum. pp.407-427, 1992.

Clore, G.L. and Storbeck, J."Affect as information about liking, efficacy, and importance." *Affect in Social Thinking and Behavior*, 2006.

Clore, G.L., Gasper, K., Garvin, E. "Affect as information." *Handbook of affect and social cognition*, 2001.

Dix, Alan J., Finlay, Janet E., Abowd, Gregory D., Beale Rusel. *Human-Computer Interaction*. Prentice Hall. 2004.

Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J., Burgelman, J-C."Scenarios for Ambient Intelligence in 2010." *ISTAG Report*, European Commission. 2001.

Dybala, Pawel, Ptaszynski, Michal, Rzepka, Rafal and Araki, Kenji. "Humor Prevails! - Implementing a Joke Generator into a Conversational System", LNAI, Vol.5360, pp.214-225, Berlin-Heidelberg, 2008.

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", In *The Proceedings of the First International Workshop on Laughter in Interaction and Body Movement (LIBM '08)*, pp.46-51, Asahikawa, Japan, June 2008.

Dybala, Pawel, Ptaszynski, Michal, Rzepka, Rafal and Araki, Kenji."Subjective, But Not Worthless -Nonlinguistic Features of Chatterbot Evaluations", The 6th IJCAI Workshop on Knowledge and Reasoning in Practical

Dialogue Systems, in *Working Notes of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp.87-92, 2009.

Dybala, Pawel, Ptaszynski, Michal, Maciejewski, Jacek, Takahashi, Mizuki, Rzepka, Rafal and Araki, Kenji. "Multiagent system for joke generation: Humor and emotions combined in human-agent conversation", Thematic Issue on Computational Modeling of Human-Oriented Knowledge within Ambient Intelligence of *The IOS Journal on Ambient Intelligence and Smart Environments*, to appear in January 2010.

Grefenstette, A.G., Qu, Y., Shanahan, J.G., Evans, D.A. "Coupling Niche Browsers and Affect Analysis for an Opinion Mining." In *Proceedings of RIAO-04*, 2004.

Hager, J. C., Ekman, P., Friesen, W. V. *Facial Action Coding System*. Salt Lake City, UT: A Human Face. 2002

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*, Vol.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", In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, pp.382-390, Honolulu, USA, October 2008.

Huitt, W. "The affective system". *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University, 2003.

Isomura, Naoki, Toriumi, Fujio, Ishii, Ken'ichiro. "Evaluation method of Non-task-oriented dialogue system by HMM", *IEICE Technical Report*, Vol.106, No.300, pp.57-62, 2006.

Jiao, Jianxin, Xu, Qianli, Du, Jun. "Affective Human Factors Design with Ambient Intelligence." In *Proceedings of The First International Workshop on Human Aspects in Ambient Intelligence*, pp.45-58, 2007.

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

Kennedy, A. and Inkpen, D. "Sentiment classification of movie and product reviews using contextual valence shifters", *Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations (FINEXIN-2005)*.

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

Landis, J.R. and Koch, G. G."The measurement of observer agreement for categorical data" *Biometrics*. Vol.33, pp.159-174, 1977.

Litman, Diane J., Pan, Shimei, Walker, Marilyn A. "Evaluating response strategies in a Web-based spoken dialogue agent", In *Proceedings of the 17th International Conference on Computational linguistics*, pp.780-786, 1998.

Loewenstein, Gorge and Lerner, Jennifer S."The Role of Affect in Decision Making." *Hanbook of Affective Sciences*, pp.619-642, 2003.

Mandel, David. Counterfactuals, emotions, and context. Cognition & Emotion, Vol.17, No.1, pp.139-159, 2003.

Miyoshi, T. and Nakagami, Y. "Sentiment classification of customer reviews on electric products", *IEEE International Conference on Systems, Man and Cybernetics*.

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

Oshima-Takane, Yuriko and MacWhinney, Brian (Ed.), Shirai, Hidetoshi, Miyata, Susanne and Naka, Norio (Rev.). *CHILDES Manual for Japanese*. McGill University, The JCHAT Project. 1995-1998.

Pang, Bo and Lee, Lillian. "Opinion Mining and Sentiment Analysis", In *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2, pp.1-135, 2008.

Polanyi, L. and Zaenen, A. (2004) 'Contextual Valence Shifters', *AAAI Spring Symposium on Exploring Attitude and Affect in Text : Theories and Applications*.

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." In *Proceedings of The 14th Annual Meeting of The Association for NLP*, pp.171-174, 2008.

Ptaszynski, Michal, Dybala, Pawel, Higuchi, Shinsuke, Rzepka, Rafal and Araki, Kenji."Affectas-Information Approach to a Sentiment Analysis Based Evaluation of Conversational Agents" In *Proceedings of the 2008 International Conference on Intelligent Agents, Web Technologies & Internet Commerce (IAWTIC'08)*, pp.896-901, Vienna, Austria, December 2008.

Ptaszynski, Michal, Dybala, Pawel, Rzepka, Rafal and Araki, Kenji."Contextual Valence Shifters Supporting Affect Analysis of Utterances in Japanese" In *Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP2009)*, pp.825-828, 2009.

Ptaszynski, Michal, Dybala, Pawel, Rzepka, Rafal and Araki, Kenji. "Affecting Corpora : Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -", In *Proceedings of The Conference of the Pacific Association for Computational Linguistics 2009 (PACLING-09)*, pp.223-228, Hokkaido University, Sapporo, Japan, September 1-4, 2009.

Ptaszynski, Michal, Dybala, Pawel, Shi, Wenhan, Rzepka, Rafal and Araki, Kenji."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)*, pp.1469-1474, Pasadena, California, USA, 2009.

Ptaszynski, Michal, Dybala, Pawel, Shi, Wenhan, Rzepka, Rafal and Araki, Kenji. "Ideas for Using Large-Scale

Corpora to Improve Verification of Emotion Appropriateness in Japanese", In *Proceedings of The Joint GCOE Symposium for Cultivating Young Researchers*, Hokkaido University, Sapporo, Japan, September 30 October 1, 2009.

Ptaszynski, Michal."Boisterous language. Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum -2channel-. (in Japanese)."M.A. Dissertation, UAM, Poznan. 2006.

Russell, James A. "A circumplex model of affect." *Journal of Personality and Social Psychology*, Vol.39, No.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." In *Proceedings of 2nd Workshop on Artificial Intelligence Techniques for Ambient Intelligence*, IJCAI＇07. 2007.

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

Schwarz, N. and Clore, G. L. "Mood, misattribution, and judgments of well－being : Informative and directive functions of affective states." *Journal of Personality and Social Psychology*, Vol.45, pp.513－523, 1983.

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

Solomon, R. C. *The Passions : Emotions and the Meaning of Life*, Hackett Publishing. 1993

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

Teixeira, J., Vinhas, V., Oliveira, E. and Reis, L. "A New Approach to Emotion Assessment Based on Biometric Data", 2nd International Workshop on Human Aspects in Ambient Intelligence (HAI＇08), In *Proceedings of the WI-IAT＇08*, pp.459－500, 2008.

Treur, Jan. "On Human Aspects in Ambient Intelligence." In Proceedings of *The First International Workshop on Human Aspects in Ambient Intelligence*, pp.5－10, 2007.

Tsuchiya, S., Yoshimura, E., Watabe, H. and Kawaoka, T. "The Method of the Emotion Judgement Based on an Association Mechanism", *Journal of Natural Language Processing*, Vol.14, No.3, pp.219－238, 2007.

Tsuchiya, Naoko. "Taiwa ni okeru kandoshi, iiyodomi no togoteki seishitsu ni tsuite no kosatsu [Statistical observations of interjections and faltering in discourse] (in Japanese)." SIGSLUD－9903－11. 1999.

Turney, Peter D. "Thumbs up or thumbs down? : semantic orientation applied to unsupervised classification of reviews." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp.417－424, 2002.

VandenBos, Gary. "APA Dictionary of Psychology". Washington, DC : American Psychological Association, 2006.

Walker, Marilyn A., Litman, Diane J., Kamm, Candace A., Abella, Alicia. "PARADISE : a framework for evaluating spoken dialogue agents", In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pp.271－280, 1997.

Yoshihira, K., Takeda, T., Sekine, S."KWIC system for Web Documents"(in Japanese). In *Proceedings of The 10th Annual Meeting of the Japanese Association for NLP*, pp.137－139, 2004.

Yu, Chen, Aoki, Paul M., Woodruff, Allison."Detecting User Engagement in Everyday Conversations"In *Proc. 8th Int'l Conf. on Spoken Language Processing (ICSLP)*, Vol.2, pp.1329－1332. Jeju Island, Republic of Korea, Oct. 2004.

Yip, Jeremy A. and Martin, Rod A. "Sense of humor, emotional intelligence, and social competence", *Journal of Research in Personality*, Vol.40, No.6, pp.1202－1208, 2006.

（2009年 6 月16日　受付）
（2009年12月 6 日　採録）

[Contact Address]
Kita-Ku, Kita 14 Nishi 9, 060-0814, Sapporo, Japan
Language Media Laboratory, 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 the Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interests include Natural Language Processing, Dialogue Processing, Affect Analysis, Sentiment Analysis, HCI and Information Retrieval. He is a member of the IEEE, SOFT, JSAI and NLP.

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 the Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interests include Natural Language Processing, Web Mining, Commonsense Retrieval, Dialogue Processing, Language Acquisition, Affect and Sentiment Analysis. He is a member of the AAAI, ACL, JSAI, IPSJ, IEICE, JCSS 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 the Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interests include Natural Language Processing, Dialogue Processing, Humor Processing, HCI and Information Retrieval.

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 research interests include Natural Language Processing, Spoken Dialogue Processing, Machine Translation and Language Acquisition. He is a member of the AAAI, IEEE, JSAI, IPSJ, IEICE and JCSS.