Affect-as-Information Approach to a Sentiment Analysis Based Evaluation of Conversational Agents

Michal Ptaszynski Pawel Dybala Shinsuke Higuchi Rafal Rzepka Kenji Araki
Graduate School of Information Science and Technology, Hokkaido University
Kita-ku, Kita 14 Nishi 9, 060-0814 Sapporo, Japan
{ptaszynski, paweldybala, shin_h, kabura, araki}@media.eng.hokudai.ac.jp

Abstract

In this paper we propose a novel method for automatic evaluation of conversational agents. The method is based on analyzing the user's affect conveyed in utterances. From analyzing: the user's general emotional engagement in the conversation and the emotion types conveyed by the user in the conversation, a simple psychological reasoning is derived about the user's sentiment about the agent's performance. The evaluation experiment on two Japanese-speaking conversational agents showed the same tendencies in the results returned by the system constructed on the proposed method and the user's opinion about the two agents checked in the afterward survey. Thus the method can be used for evaluation of Japanese-speaking conversational agents.

1. Introduction

Technological development led to creating a new dimension of communication, where a machine is an object, a Human-Computer Interaction (HCI) [1]. What was carried whit it, was a rush in development of intelligent talking agents, like car navigation systems [2] or talking furniture [3]. Their functional implementation into our everyday lives has already become a current process, and a need for humanized HCI grows rapidly.

Along with it people found themselves in a need for developing a quick, automatic evaluation method for such agents. The usual methods used are asking the opinions of users-testers about their satisfaction while using an agent, the naturalness of agent's utterance producing, the will for continuing the conversation, etc.

However, since decision-making and therefore expressing opinions in humans strongly depend on such features, as their emotional states, imagination and experiences [4], such opinions are always stigmatized with a lack of objectivity. Therefore we propose a novel method of evaluation for Japanese speaking conversational agents based on classifying users' sentiment towards an agent. The sentiment classification is based on a sociopsychological reasoning of "affect as information" [5] derived from affect analysis of the user's utternaces.

2. Our approach – affect derived sentiment

2.1. Sentiment analysis for agent evaluation

Sentiment analysis is a sub-topic of information extraction that only recently has gained interest of scientists [6]. The general idea of sentiment analysis, which is to gather and classify attitudes (into positive or negative) about particular topics or entities, is important for marketing research, monitoring of chat-rooms' content for security reasons [7], or customer feedback on particular products.

In the case of agents, it is desirable to acquire objective information about their performance before putting them on the market, since a failure might bring losses of funds and human effort. Unfortunately, tests, where people are hired to verify the usability and performance of market-destined agents, are burdened with a lack of objectivity from the very fact of paying people money for the evaluation. There is no other way of performing such test, than making a human talk to an agent; but there is a method of gathering objective information for evaluation of the product then a typical survey – where no survey is needed at all.

In our assumption, gathering information about the tester's sentiment towards the product during the test will provide objective evaluative information. In our method, we perform this by analyzing the affect of user's utterances and deriving from it the user's sentiment to the agent interlocutor.

2.2. Analyzing affect for sentiment estimation

Affect analysis is a sub-field of information extraction. Its goal is to estimate human emotional states in different kinds of communication. Popular methods for analyzing affect include analyzing emotions from facial expressions or voice [8]. However, since emotions are strongly context dependent [21], most of the semantic content of expressing emotions is ignored in such researches. Therefore we decided to use a method to analyze the affect of a textual input for the need of usage in HCI.

There are several researches on affect analysis [9]. However there have been only few approaches to apply the affect analysis to gather information about sentiment and attitudes [10] and no significant work was done on applying such approach to evaluation of conversational agents talking in Japanese.

2.3. Affect as information

The notion of "affect as information" was introduced in 1983 by Schwarz and Clore [5] and is widely studied in the field of psychology and social psychology. Schwarz and Clore claimed that people use affect just as any other criterion, by applying the informational value of their affective reactions to form their judgments, attitudes and opinions. For example, when we talk to someone whose behavior evokes in us only negative emotions, our attitude to such person will be rather negative. This results in our feedback behavior, like distancing from that person or lack of a will to continue the conversation.

Applying this thought to evaluation of conversational agents should indicate similar tendencies in the results of affect analysis-derived sentiment classification and the results acquired through an objective and non-commercial survey.

2.4. Our goal

If this appeared to be true, the method in its final application would be capable to substitute the laborious and uneconomical surveys. Today's evaluation methods bring along costs of preparing and carrying out the survey. Not mentioning the costs of printing the suvery sheets, the money paid to the users-testers is high because of sophisticated preparations, and the responsibility to judge the product properly.

However, if we only sit the users before an interface, let them talk freely and gather the evaluative information from their behavior during the conversation, the problem of preparing and printing the survey will disappear. Moreover, the evaluation itself will not be burden with influential responsibility or a possibility that the user's attitude to the product would change for some reason in the time between the end of the conversation and filling out the survey.

However, for the first step of development of such evaluation method, we want to find out, whether similar tendencies appear in a non-commercial survey and the proposed method. If they did, the method would be destined for further development, and in its present shape usable as a preliminary evaluation method for more detailed verification of agents' performance.

3. ML-Ask for affect analysis

As the affect analysis system we decided to use Ptaszynski's at al. Emotive Elements / Emotive Expressions Analysis System (ML-Ask) [9]. The system is designed to analyze affect from textual input in the Japanese. The process of analysis in the system 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 are the emotions conveyed in the utterance, or setting the emotive value; 3. If the sentence is emotive, finding the types of emotions conveyed in the utterance.

The method of analyzing the affect is based on Ptaszynski's [11] idea of binary classification of realizations of emotions in language into:

1. *Emotive elements*. Informing that emotions have been conveyed, but do not expressing specified feelings or expressing different ones, depending on the context. Examples are: *sugee* (great!), *waku-waku* (heart pounding), *-yagaru* (fu**ing do sth);
2. *Emotive expressions*. Parts of speech, which, in emotive sentences, describe emotional states. Examples are: *aijou* (love), *kanashimu* (feel sad), *ureshii* (happy).

In a textual input provided by the user three features (emotiveness, emotive value and emotional state) are determined by cross-referencing the databases of emotive elements with emotive expressions. The emotive elements databases were gathered from different research [12, 13, 14, 15] and divided into interjections, mimetics, endearments, vulgarities and representations of non-verbal emotive elements, like exclamation marks or ellipsis. Also there was added an algorithm detecting emoticons, as they are symbols commonly used in everyday text-based communication tools. The databases of emotive expressions were created on Nakamura's [16] collection.

3.1 ML-Ask's output as information

In our approach we use ML-Ask to analyze utterances of a user talking to a conversational agent. The results of the analysis are then viewed as follows. First, if many [1] of the user's utterances were determined as emotive, this means that the user was emotionally involved in the conversation. Emotional involvement means for the user a tendency to easier familiarizing with the interlocutor, and along with it, loosing the sense of identifying a machine in it, ergo considering the machine as more human. Second, analysis of valence of specified emotion types conveyed by the user in the whole conversation provides us information on what were the user's feelings towards the machine interlocutor during the conversation. If the feelings were positive, or changing on a vector negative→(neutral)→positive while talking, the attitude, and therefore the general sentiment towards the agent is considered as positive. If the emotions were negative or changing on a vector positive→(neutral)→negative, the sentiment is classified as negative.

Both types of information acquired provide a wide overview on the user's sentiment about the agent and, although being able to be analysed separately, it is desirable for the both types of information to harmonize rather then show dissonance.

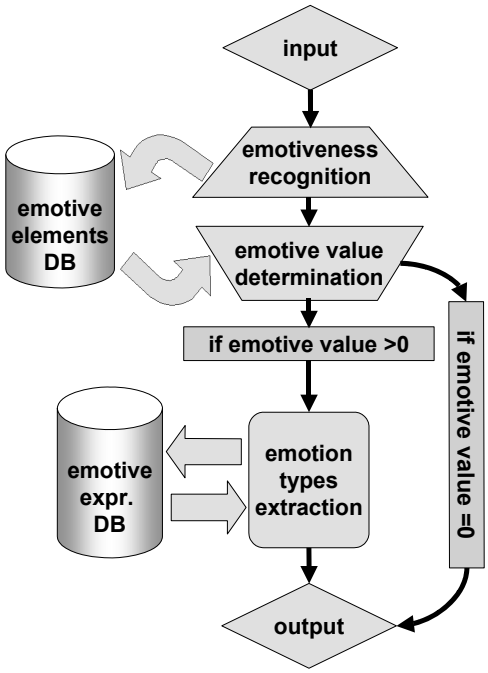

Fig. 1. Flow chart of ML-Ask.

[1] We do not precise the value of "many". Instead of that we compare the results for two different conversational agents.

4. Applying 2-dimensional model of affect

The idea of 2-dimensional model of affect was first proposed by Schlosberg [17] and developed further by Russell [18]. Its main assumption is that all emotions can be described in a space of two-dimensions: the emotions' 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.

4.1. Classification of emotions

Nakamura [16], after a thorough study on emotions in the Japanese, proposes a classification of emotions into 10 types - said to be the most appropriate for the language. That is: *ki* / *yorokobi* (joy, delight), *do* / *ikari* (anger), *ai* / *aware* (sorrow, sad-ness), *fu* / *kowagari* (fear), *chi* / *haji* (shame, shyness, bashfulness), *kou* / *suki* (liking, fondness), *iya* / *iyodomi* (dislike, detestation), *kou* / *takaburi* (excitement), *an* / *yasuki* (relief) and *kyou* / *odoroki* (surprise, amazement).

4.2 Nakamura's emotions in two dimensions

We grouped the emotion types distinguished by Nakamura on the Russell's model (see Fig. 2). The types fitting in both, like *kou* / *takaburi*, which contains expressions of positive "excitement" as well as negative "irritation", were put in both possible quarters. However none of the types was placed in more than two quarters.

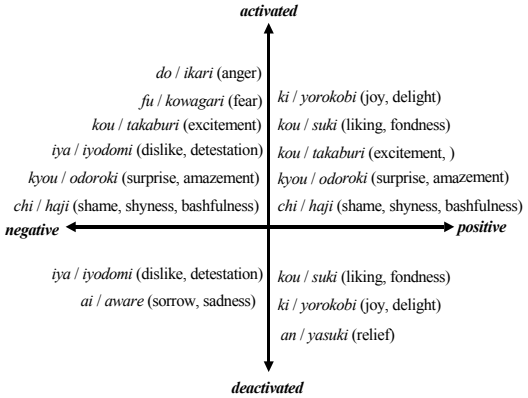

Fig. 2. Grouping Nakamura's classification of emotions on Russell's two-dimensional space.

5. Evaluation experiment

To test the method we performed an evaluation of two conversational agents. We asked five students (1 female, 4 males) to perform a 10-turn conversation with both agents. No topic restrictions were made, so

that the talk could be as free and human-like as possible. The agents were first evaluated during the conversation using the proposed method. After the conversation the users were asked to fulfill a questionnaire concerning their sentiment about the agents. The results acquired by the method and provided by the users for the two agents were compared looking for similarities in sentiment classification.

5.1 Two agents – a short description

5.1.1. Modalin. A non-task oriented keyword-based conversational agent, which uses modality to enhance Web-based propositions for dialogue. The agent was developed by Higuchi et al. [19].

5.1.2. Pundalin. A non-task oriented conversational agent created by combining Modalin with Dybala's Pun generating system PUNDA [20]. Pundalin therefore is a humor-equipped conversational agent using puns to enhance the communication with a user.

5.2. Questions we asked users and their representations in sentiment analysis

5.2.1. User's evaluation. The questions asked were: A) Do you want to continue the dialogue?; B) Was the agent's talk grammatically natural?; C) Was the agent's talk 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 agent's talk interesting and funny?. The answers for questions were given in 5-point scale 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?"

5.2.2. Representations in sentiment analysis. We made the following assumptions on how the questions we asked users directly were represented by the results provided by the analysis. The questions A)–H) generally illustrate how much users could familiarize with the agent, involve emotionally in the conversation and find the machine more human-like. Therefore all specific questions represented the first type of information, and the general summarizing question was directly corresponding to the second type of information acquired from the sentiment analysis – valence polarity of emotion types.

6. Results

The results of the evaluation are showed below. First the results of the questionnaire are shown and after that the results of the analysis are summarized and compared to the users' opinions contained in the survey.

6.1 User's evaluation

4 out of 5 users evaluated Pundalin (humor-equipped agent) as better than Modalin (see Fig. 3). Pundalin received higher scores also in detailed questions (see Fig. 4.). The difference was especially visible in questions B, G and H.

Although not in all categories the differences between Modalin and Pundalin were that significant, overall results for both agents clearly showed, the performance of which was estimated as a more human-like and easy to familiarize.

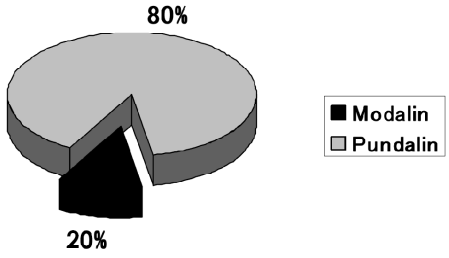

Fig. 3. User's evaluation-results for the question "Which agent do you think was better?"

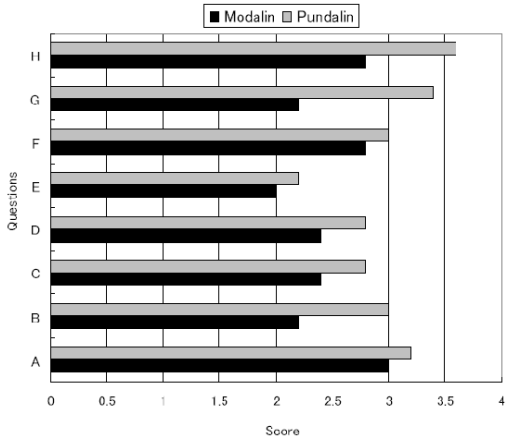

Fig. 4. User's evaluation - results for Modalin and Pundalin for detailed questions (see 5.2.1.). Answers were given in a 5-point scale.

6.2. Results of sentiment analysis

Evaluation based on sentiment analysis of the users' utterances showed tendencies similar to the survey evaluation. The users were more emotionally involved in the conversations with Pundalin, which in our

assumption corresponded to the direct opinions about the agent's performance - that it was more human like, its utterances were more correct semantically and grammatically, etc. (see 5.2.1. and 6.1., Fig. 5.).

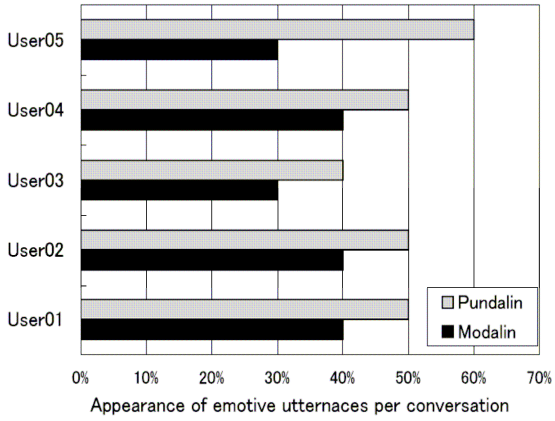

Fig. 5. Percentage of average appearance of emotively engaged utterances for all five users in conversations with both agents.

The analysis of specified emotion types conveyed by the users in conversations provided the information clearly showing the users' feelings and attitudes towards the both agents. The users' general attitudes were in 80% positive for Pundalin whereas for Modalin the attitudes of the users were only negative (see Fig. 6.).

Therefore we can say that the general sentiment of a user towards an agent was positive for Pundalin and negative for Modalin.

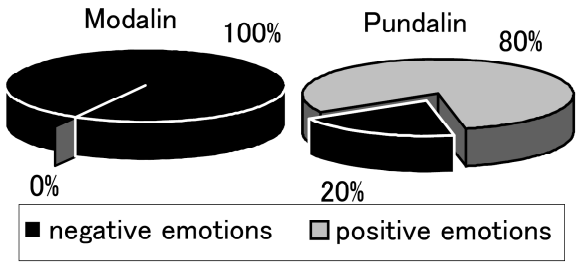

Fig. 6 The total relation of emotions positive to negative conveyed in the utterances of users with Modalin and Pundalin.

7. Conclusions

In this paper we presented a novel method of evaluation for conversational agents. The method is based on analyzing affect conveyed by a user in a conversation with an agent. Borrowing the notion of "affect as information" [5], in the results of affect analysis performed by a system created by Ptaszynski et al. [9] provide us information about the user's emotional involvement in a conversation, loosing a psychological distance, and easiness in familiarizing with the machine. This corresponds to the direct questioning the user about the agent's performance. Furthermore, analysis of specified emotion types conveyed by the user in the whole conversation and their classification by applying the two-dimensional model of emotions [18] provides us information on what were the user's attitudes towards the machine interlocutor during the conversation.

In the evaluation experiment performed on two conversational agents there have been seen similar tendencies in the results acquired by the method and the results of the questionnaire with direct questions to the users about the agent's performance. Comparing to the traditional user-oriented surveys, the proposed method is non-invasive and can provide objective information on user's sentiment to machine-interlocutor on the spot. Since an approximate time of processing one utterance is 0.143 s, the method can be used in real time, and provide actual information on changes in the user's attitude towards the machine. This feature does not only provide a faster and more up-to-date information on the user's sentiment, but also, appropriately utilized, can provide hints for the agent about the potential undesirable changes in the user's attitudes and the need for appropriate counteractions, during an everyday use.

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 an evaluative mean, the method saves time, effort and funds spend each time on preparing and performing laborious surveys.

The method is still not perfect. For now it shows only the general tendencies in the users' attitudes to agents in a simple comparison of two (or more) agents. We will continue developing the method to work out more precise countable units for evaluation.

However, in the first step of development our goal was to check whether there were similar tendencies in a survey evaluation and the proposed method. Since such tendencies were confirmed very clearly, the method is generally applicable and in its present shape usable as a preliminary evaluation method for more detailed verification of agents' performance or a strong supportive mean to objectivize the results of traditional questionnaires.

8. Discussion and future work

The method, although proved to be effective, has still some lacks, desirable to be filled in the near future. The imperfections of the sub-systems used in the method influence its accuracy. The slight lacks in emotion types extraction procedure in ML-Ask limit

the information about emotional states conveyed by the user in a conversation. Also some defects of tools for morphological analysis of utterances used in ML-Ask decrease the system's performance. However, it is predictable that using the two-dimensional model of emotions [18] to specify the emotional affiliations of emotive elements will disambiguate the databases and therefore improve the performance of emotion types extraction in ML-Ask.

Moreover, the notion of "affect as information", although with a firm scientific background in psychology and social psychology [22, 23], is not a common notion in the fields we referred to in this paper – agent development, evaluation methods development, affect analysis and so on. Coordinating the appropriate items to evaluate automatically with the questions asked directly to the user is based on psychological reasoning, and therefore reaches deeper and beyond the simple numbers usually put it in terms of so familiar notions of precision or recall. However, the rapid development in all fields of science, as well as in commercial areas, makes researchers from different scientific fields join the efforts – as we shown in this paper – successfully.

Acknowledgements

This work was partially supported by the Research Grant from Nissan Science Foundation.

References

[1] A.J. Dix, A. Dix, J. Finlay, G.D. Abowd. *Human-Computer Interaction*. Prentice Hall. 2004.

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

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

[4] Rafal Rzepka, and Kenji Araki. "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.

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

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

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

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

[9] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, and Kenji Araki. "Effective Analysis of Emotive-ness in Utterances Based on Features of Lexical and Non-Lexical Layers of Speech." *Proceedings of The 14th Annual Meeting of The Association for NLP*. 2008. pp. 171-174.

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

[11] Michal Ptaszynski. "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.

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

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

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

[15] Jonas Sjobergh. *Vulgarities are fucking funny, or at least make things a little bit funnier*. Proceedings of KTH CSC, Stockholm. 2006.

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

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

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

[19] Shinsuke Higuchi, Rafal Rzepka, and Kenji Araki. "Web wo riyoshita renso tango oyobi modarithi-hyougen ni yoru zatsudan shisutemu [Chat system based on modality expressions and association words extracted form the Web] (in Japanese)." *Proceedings of The 14th Annual Meeting of The Association for NLP*. 2008. pp. 175-178.

[20] Pawel Dybala, Michal Ptaszynski, Rafal Rzepka, and Kenji Araki. "Extracting Dajare Candidates from the Web - Japanese Puns Generating System as a Part of Humor Processing Research." *Proceedings of International Workshops on Laughter in Interaction and Body Movement (LIBM'08)*. 2008. pp. 46-51.

[21] David Mandel. Counterfactuals, emotions, and context. *Cognition & Emotion*, 17(1). 2003. pp. 139-159.

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

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