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

Michal PTASZYNSKI
ptaszynski@media.eng.hokudai.ac.jp

Pawel DYBALA Shinsuke HIGUCHI Rafal RZEPKA Kenji ARAKI

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

Presentation Outline

1. Introduction
2. Assumptions for Evaluation:
   - Sentiment Analysis
   - Affect as Information
3. Affect Analysis System
4. Experiment
5. Results
6. Conclusions and Future Work

Kinds of Agents

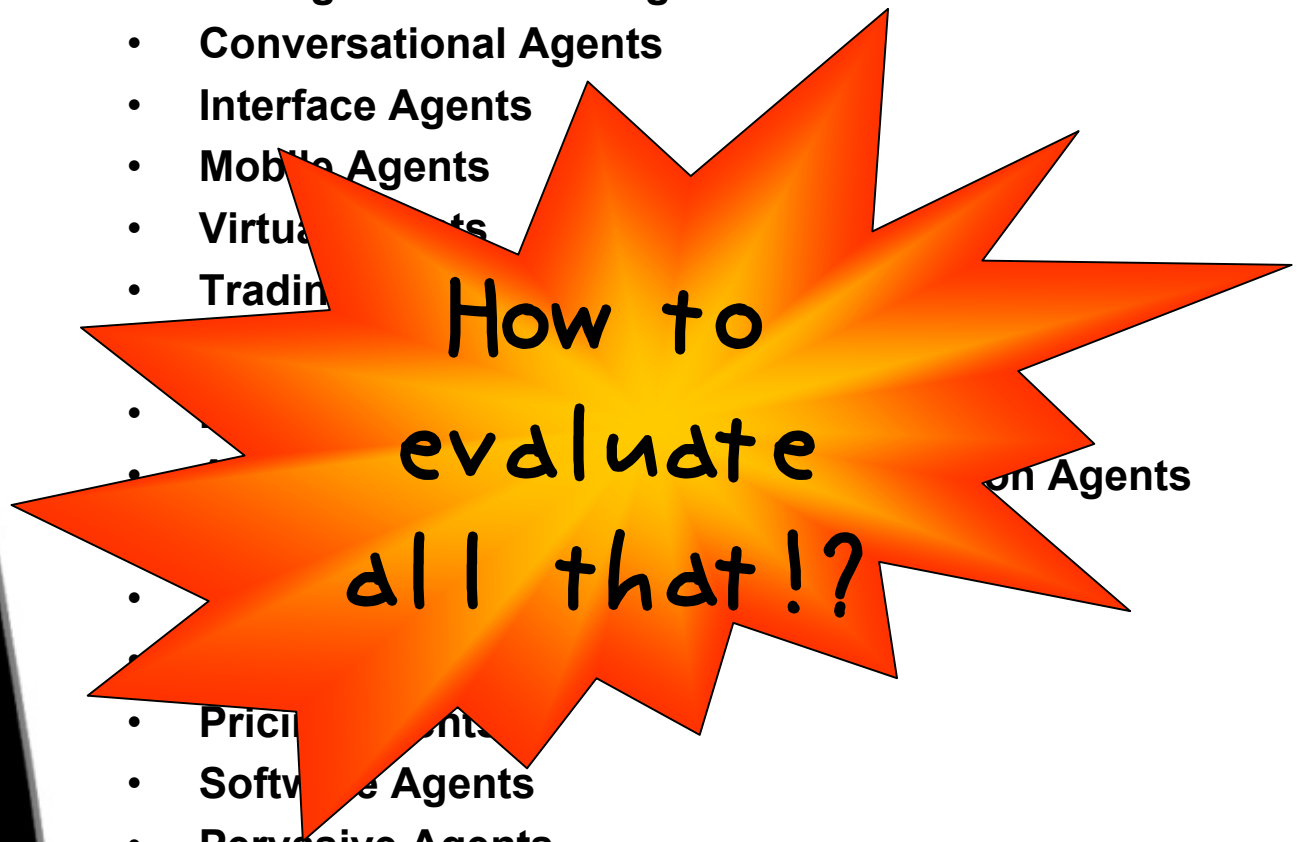
…apart from agent 007

- **Intelligent Business Agents**
- **Conversational Agents**
- **Interface Agents**
- **Mobile Agents**
- **Virtual Agents**
- **Trading Agents**
- **E-commerce Agents**
- **Believable Agents**
- **Autonomous Knowledge and information Agents**
- **Information Filtering Agents**
- **Task-Oriented Agents**
- **Mediating Agents**
- **Pricing Agents**
- **Software Agents**
- **Pervasive Agents**
- **Ubiquitous Systems and E-Technology Agents**
- **...etc.**

Kinds of Agents

…apart from agent 007

- **Intelligent Business Agents**
- **Conversational Agents**
- **Interface Agents**
- **Mob[illegible] Agents**
- **Virtu[illegible]**
- **Tradi[illegible]**
- **[illegible]**
- **[illegible]on Agents**
- **[illegible]**
- **[illegible]**
- **Prici[illegible]**
- **Softw[illegible] Agents**
- **Pervasive Agents**
- **Ubiquitous Systems and E-Technology Agents**
- **...etc.**



A Bit of Statistics

CIMCA / IAWTIC / ISE Conferences:

- About 160 papers (many on agent technology)
- 21 papers with word « Agent»
  - **Only few (3) on evaluation methods for agents...**

...Why so few?

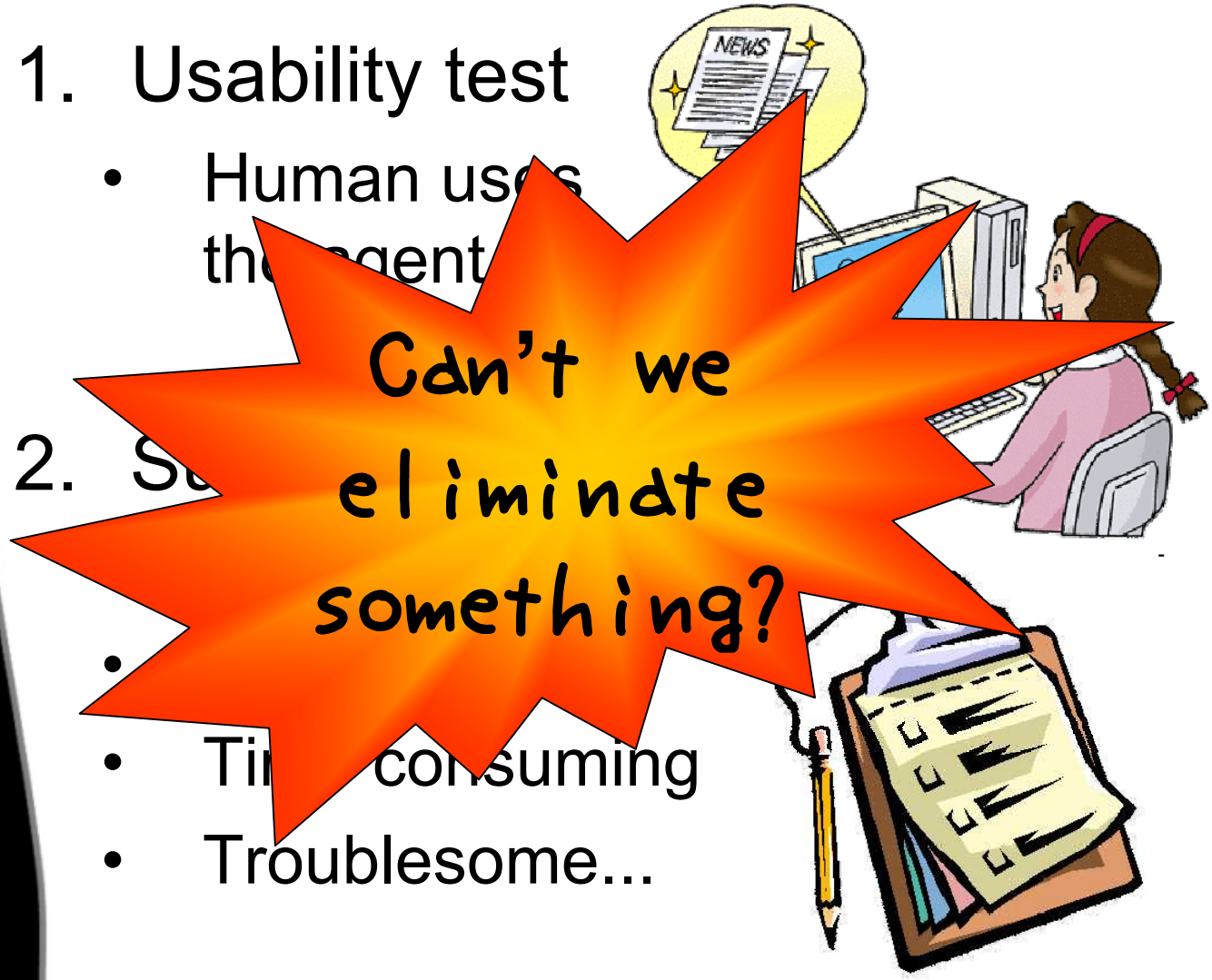
Common evaluation methods

1. Usability test
   - Human uses the agent

2. Survey
   - Subjective
   - Expensive
   - Time consuming
   - Troublesome...

Common evaluation methods

1. Usability test
   - Human use[illegible] th[illegible]gent
2. S[illegible]
   - [illegible]
   - Ti[illegible] consuming
   - Troublesome...



Common evaluation methods

1. Usability test
   - Human uses the agent

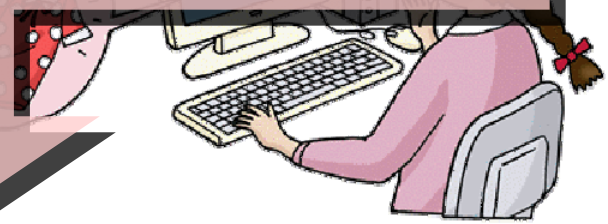
2. Survey
   - Subjective
   - Expensive
   - Time consuming
   - Troublesome...

Cannot eliminate that ...yet

Common evaluation methods

1. Usability test
   - Human uses the agent

Maybe we could get rid of this?



2. Survey
   - Subjective
   - Expensive
   - Time consuming
   - Troublesome...

Kinds of Agents

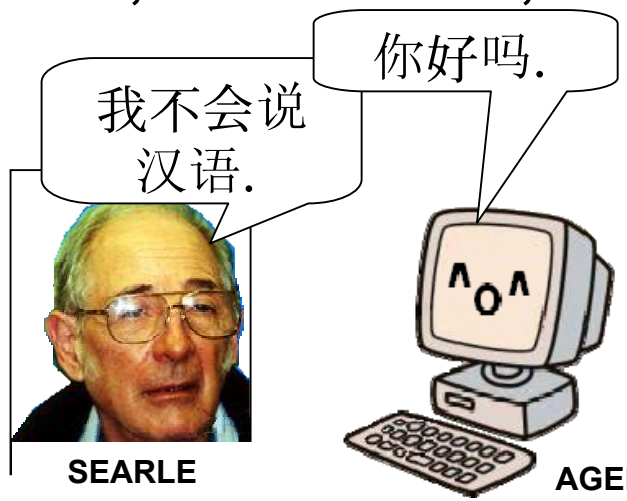
…apart from agent 007

- Intelligent Business Agents
- Conversational Agents
- Interface Agents
- Mobile Agents
- Virtual Agents
- Trading Agents
- E-commerce Agents
- Believable Agents
- Autonomous Knowledge and information Agents
- Information Filtering Agents
- Task-Oriented Agents
- Mediating Agents
- Pricing Agents
- Software Agents
- Pervasive Agents
- Ubiquitous Systems and E-Technology Agents
- ...etc.

Conversational agents

Conversational agent – communication technology that utilize natural language and computational linguistic techniques to engage users in human-like Web-based “dialogs.”

Application in: business enterprises, education, government, healthcare, entertainment, etc.



Conversational agents

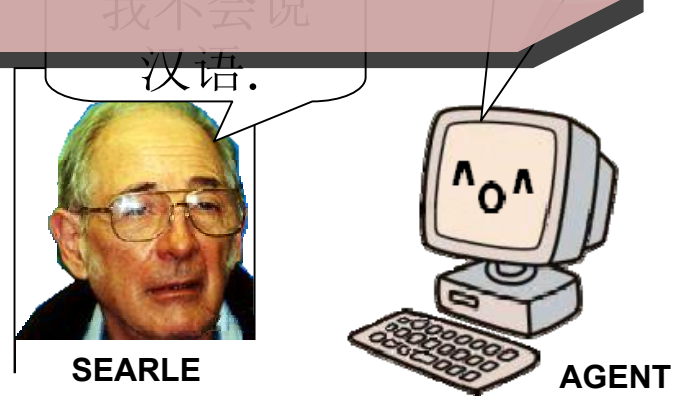
Conversational agent –

communication technology that utilize natural-language and computational linguistic techniques to engage users in human-like Web-based "dialogs."

Task:
Satisfy user through conversation

Desirable features:
1. Human-likeness
2. Arousing positive attitude in User

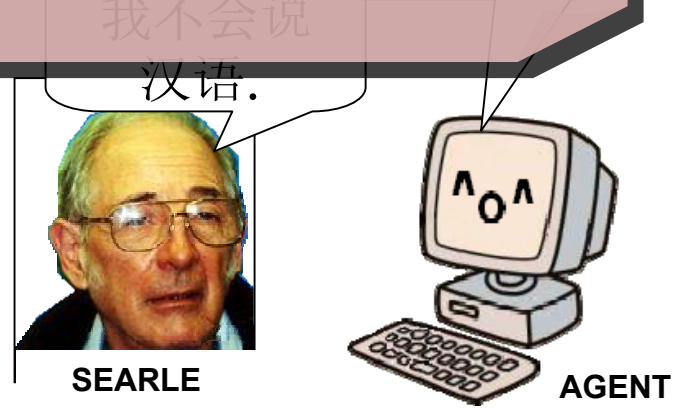
Application in: business enterprises, education, government, healthcare, e-Education, positive attitude, Internet



Can these be checked automatically?

Desirable features:

1. Human-likeness
2. Arousing positive attitude in User



Evaluation –

Slightly different approach

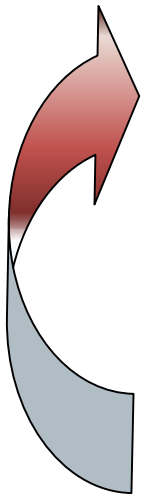
- How about using an automatic method of attitude estimation in real time during a user-agent conversation?

Evaluation –

Slightly different approach

- We propose a novel method of estimating the user's attitude and sentiment towards a Japanese-speaking agent, based on "affect as information" reasoning derived from affect analysis of the user's utterances.

Definitions



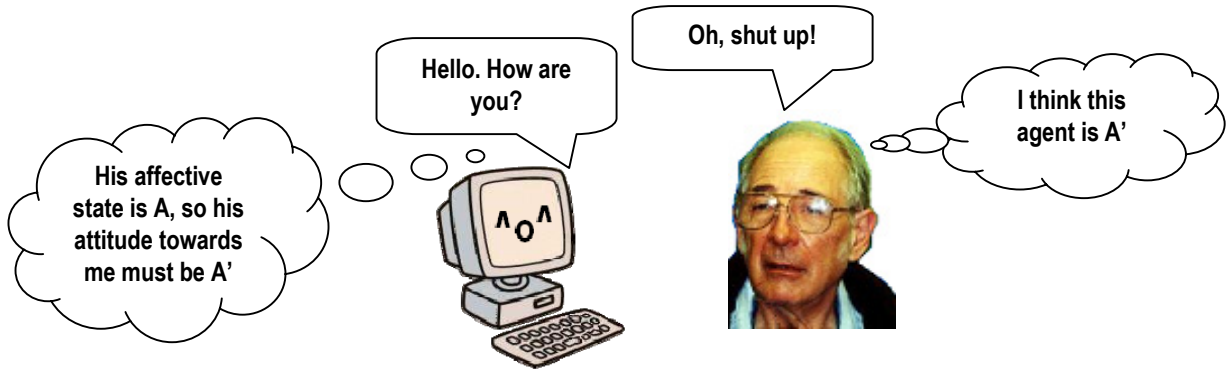
Sentiment analysis is a sub-field of information extraction aiming to estimate the attitude (e.g. positive/negative) of a speaker or a writer towards / with respect to some topic (e.g. about a product).

Affect analysis is a sub-field of information extraction. Its goal is to estimate emotional states in humans.

Affect as information

"People use affect as information as a criterion, by applying the informational value of their affective reactions to form their judgments, attitudes and opinions."

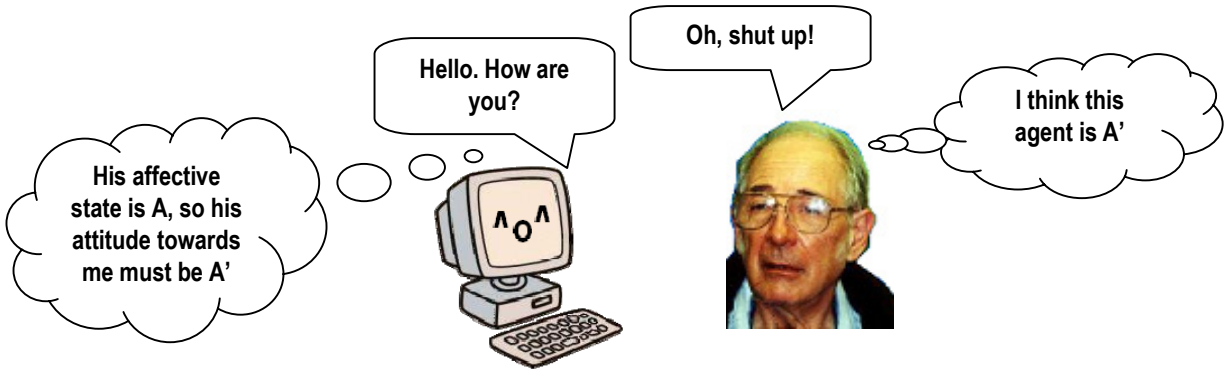
If we new affective states of a user during his conversation with an agent, we should be able to derive from it a reasoning about their judgments during filling in the survey (=attitudes/sentiment to the interlocutor). So…



• 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*, 45: 513–523. 1983.

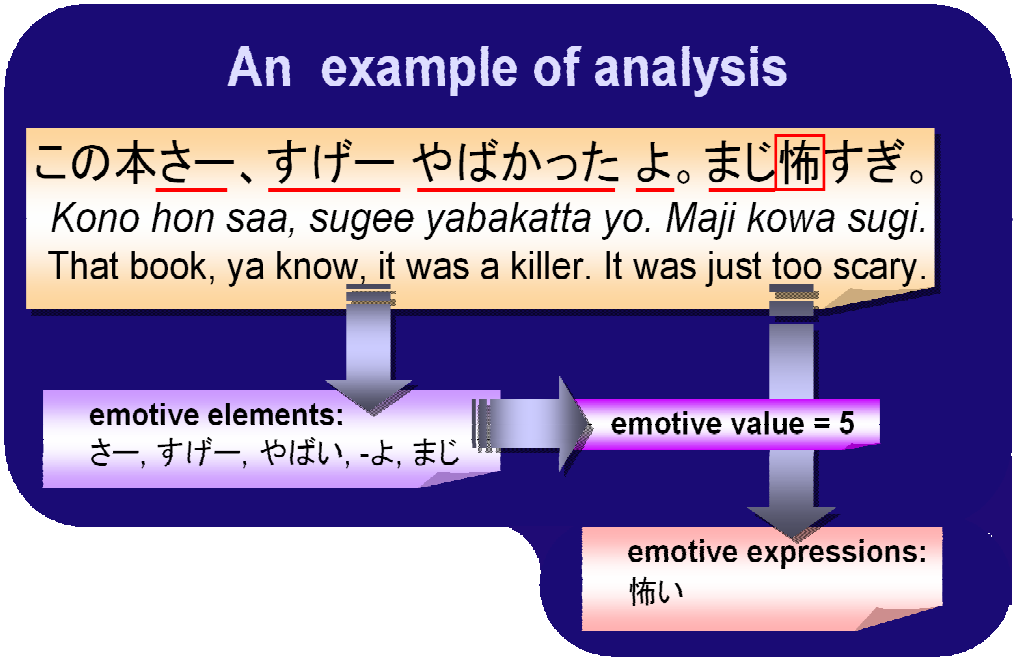
Affect as information

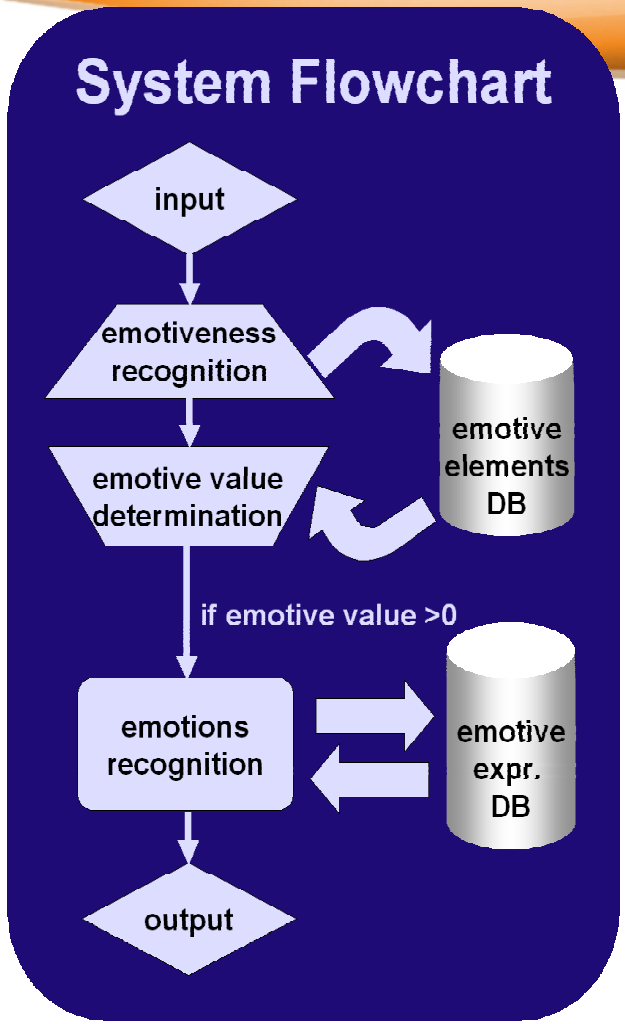
- There should be similar tendencies in the results acquired through a usual survey and the results of affect analysis-based sentiment classification.



• 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*, 45: 513–523. 1983.

ML-Ask - affect analysis system





Definition of Emotions

- Emotions = every temporary state of mind, feeling or emotional state evoked by experiencing different sensations.

- Emotive utterances = every utterance in which the speaker is emotionally involved, and this involvement is linguistically expressed.

- 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)
- Nakamura, A.: Kanjo hyogen jiten (Dictionary of Emotive Expressions) (in Japanese), Tokyodo Publishing, Tokyo (1993)

Classification of Emotions

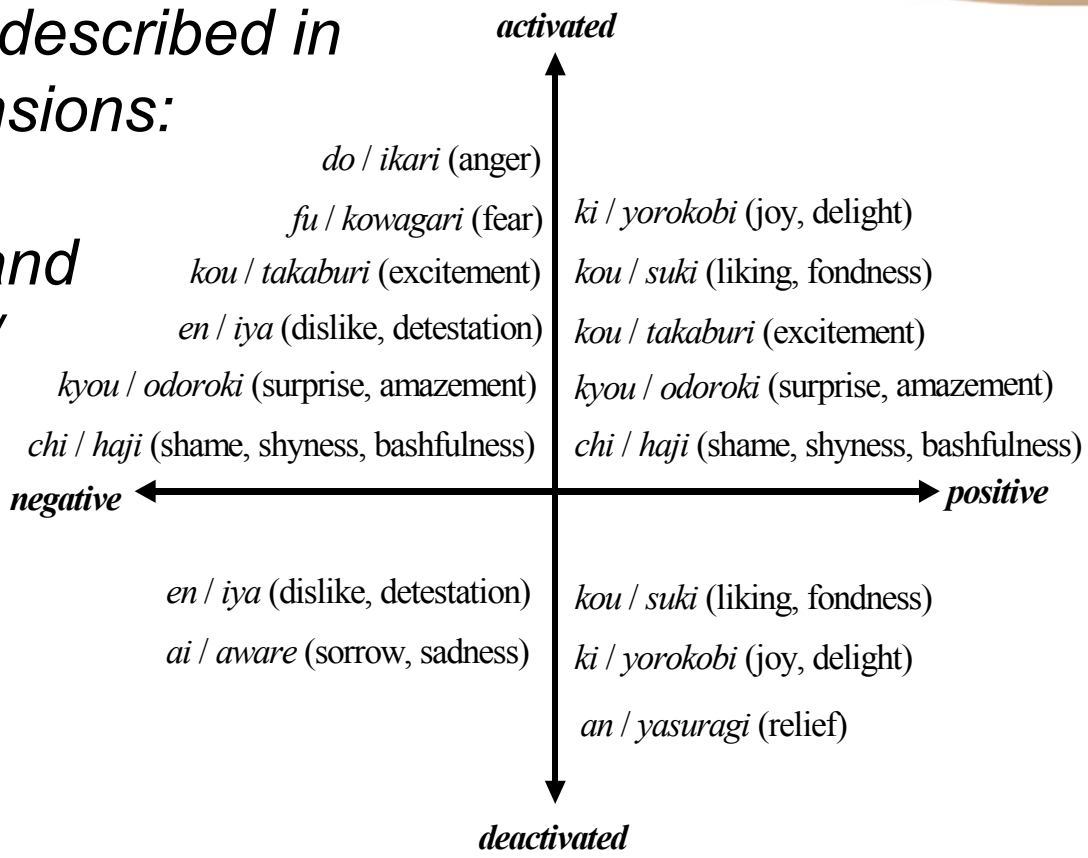
- Nakamura's classification of emotions after a thorough study in the Japanese:

  10 types said to be the most appropriate for the Japanese:

  1. 喜 *ki* / *yorokobi* (joy, delight)
  2. 怒 *do* / *ikari* (anger)
  3. 哀 *ai* / *aware* (sorrow, sadness)
  4. 怖 *fu* / *kowagari* (fear)
  5. 恥 *chi* / *haji* (shame, shyness, bashfulness)
  6. 好 *kou* / *suki* (liking, fondness)
  7. 厭 *en* / *iya* (dislike, detestation)
  8. 昂 *kou* / *takaburi* (excitement)
  9. 安 *an* / *yasuragi* (relief)
  10. 驚 *kyou* / *odoroki* (surprise, amazement)

2-dimensional model of affect

"All emotions can be described in a space of two-dimensions: valence polarity (positive / negative) and activation (activated / deactivated)."



H. Schlosberg. "The description of facial expressions in terms of two dimensions." Journal of Experimental Psychology, 44:229-237. 1952.
James A. Russell. "A circumplex model of affect." Journal of Personality and Social Psychology, 39(6):1161-1178. 1980.

ML-Ask's output as information

- ML-Ask analyses the user's utterances.

1. It classifies whether they are emotive.
   - If many utterances were emotive (comparing two agents)
     → The user was more emotionally involved in the conversation
     ⇒ In the Japanese, emotional involvement in a conversation suggests a tendency to easier familiarizing with the interlocutor
     ⇒ loosing the sense of recognizing a mere machine in the agent ＝considering the agent as more human = BETTER PERFORMANCE
2. It classifies the emotion types
   - If emotions were: [**positive**], [**neutral**]→[**positive**], [**negative**]→[**positive**], the user's ATTITUDE = POSITIVE.

Agents for evaluation

Modalin. A non-task oriented keyword-based conversational agent, which uses modality to enhance Web-based propositions for dialogue.

Pundalin. A non-task oriented conversational agent created on the base of Modalin combined with pun generating system. Pundalin therefore is a humor-equipped conversational agent using puns to enhance the communication with a user.

- 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*:175-178. 2008.
- 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)*: 46-51. 2008.

Survey

1. Detailed questions (5-point scale with explanations):

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?;
H) Did you find agent's talk interesting and funny?;

Survey

2. Summarizing question:

“Which agent do you think was better?”

Representation of questions in sentiment analysis

- B-D: how high did the users evaluate agents' talking abilities;
- A, E-F: how much could the users familiarize with agents
- A, G-H: how much could the users get involved emotionally in the conversation.
  →particular questions ⇒ OPINIONS ABOUT PERFORMANCE

Representation of questions in sentiment analysis

- The general summarizing question = POSITIVE / NEGATIVE ATTITUDES

Survey vs. affect analysis

Re 1. **Particular questions about performance vs number of emotive utterances**

Survey: Pundalin received higher scores in detailed questions.
Affect Analysis: The users were more emotionally involved in the conversations with Pundalin.

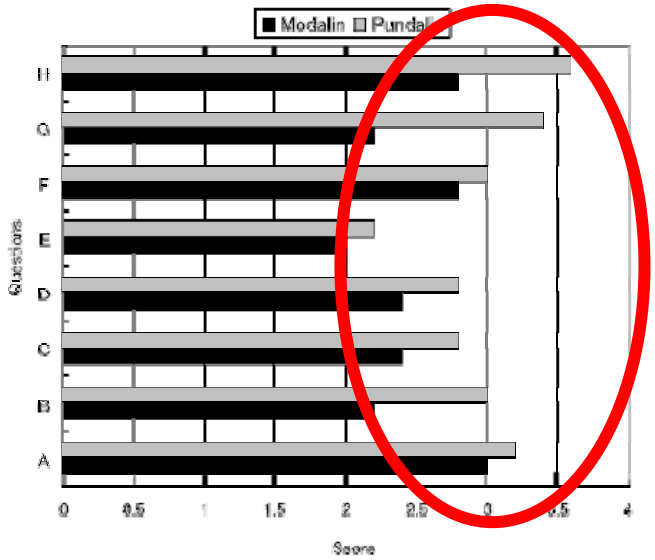


Fig. 4. User's evaluation - results for Modalin and Pundalin for detailed questions (see 5.2.1.). Answers were given in a 5-point scale.

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

Survey vs. affect analysis

Re 2. **Summarizing question about attitude**

Survey: 4 out of 5 users (80%) evaluated Pundalin (humor-equipped agent) as better than Modalin.

Affect analysis: The users' general attitudes to Pundalin were in 80% positive whereas to Modalin the attitudes of the users were only negative.

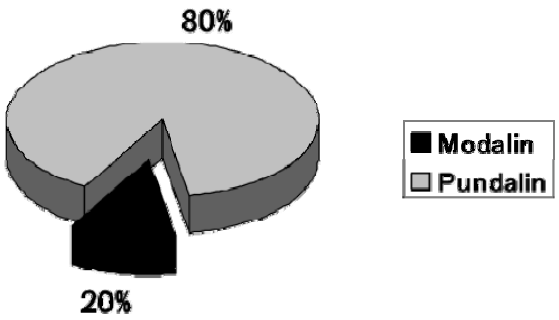


Fig. 3. User's evaluation-results for the question "Which agent do you think was better?"

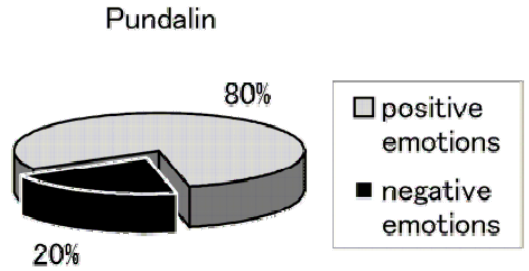


Fig. 7. The total relation of emotions positive to negative conveyed in the utterances of users with Pundalin.

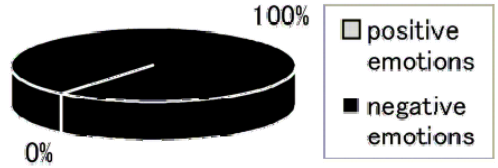


Fig. 6. The total relation of emotions positive to negative conveyed in the utterances of users with Modalin.

Conclusions

- Similar tendencies - found
- Experiment (though still rather small) – confirmed Affect-as-Information theory within HCI
- Worth continuing!
- Features of method:
  - non-invasive
  - provide objective information (on user's sentiment about machine-interlocutor)
  - can be used in real time (processing time = 0.143 s per utterance)
- As an evaluation method:
  - saves time, effort and funds spend on preparing and performing laborious surveys
- Potential applicability (more than evaluation method):
  - Provide a up-to-date information on the user's sentiment
  - Provide hints for the agent about the need for appropriate counteractions, in an everyday use.

Future Work

- **Larger number of partiocipants in evaluation**
- **Other kinds of conversational agents**
  (With jokes the similarities were vivid – how about comparing non-joking agents?)
- **Specify mapping**
  (what exactly informationcan be derived from affective states?)
- **Find universalities**
  (now – only comparison of two or more agents; in the future – evaluating also one agent)
- **Improving intermediary tools**
  (affect analysis, part of speech analyser)

Thank you for your kind attention.