International Journal on Artificial Intelligence Tools
Vol. XX, No. X (2010) 1–39

World Scientific
www.worldscientific.com

EVALUATING SUBJECTIVE ASPECTS OF HCI ON AN EXAMPLE OF A NON-TASK ORIENTED CONVERSATIONAL SYSTEM

PAWEL DYBALA*

Graduate School of Information Science and Technology
Hokkaido University, Kita 14 Nishi 9, Kita-ku
060-0814 Sapporo, Japan
paweldybala@media.eng.hokudai.ac.jp

MICHAL PTASZYNSKI

Graduate School of Information Science and Technology
Hokkaido University, Kita 14 Nishi 9, Kita-ku
060-0814 Sapporo, Japan
ptaszynski@media.eng.hokudai.ac.jp

RAFAL RZEPKA

Graduate School of Information Science and Technology
Hokkaido University, Kita 14 Nishi 9, Kita-ku
060-0814 Sapporo, Japan
kabura@media.eng.hokudai.ac.jp

KENJI ARAKI

Graduate School of Information Science and Technology
Hokkaido University, Kita 14 Nishi 9, Kita-ku
060-0814 Sapporo, Japan
araki@media.eng.hokudai.ac.jp

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

The evaluation of subjective aspects of HCI, such as human-likeness, likeability or users' emotions towards computers is still quite a neglected issue, especially in the field of non-task oriented conversational systems (chatterbots). In this paper we try to bridge this gap by proposing a new methodology of evaluation. The methods presented were tested in our research on humor-equipped chatterbots. We describe them in details, discuss their drawbacks and usability. In one of the presented methods we used an emotiveness analysis system, which itself can be considered an AI tool, as it was used to detect users' emotions towards conversational systems, and to perform their automatic evaluation. We also propose some methods that we have not used yet, which, however,

* Corresponding author:
paweldybala@media.eng.hokudai.ac.jp (Pawel Dybala)

seem applicable in this field, such as brain scanning techniques. Finally, we give some ideas that should be addressed in the future.

Keywords: HCI, conversational systems, evaluation methodology

1. Introduction

Evaluation of such vague and hard to define issues as "interaction" or "conversation" (both between humans and in HCI) is generally quite a difficult task. There are some features that can be assessed relatively objectively, such as participant's communication skills, linguistic skills or understandability, and a detailed analysis may give us some actual numbers here (amount of mistakes, misunderstood sentences etc.). However, there are some more troublesome issues to evaluate – looking from an interaction participant point of view, we can assess partner's likeability, pleasure we had from the interaction, partner's funniness, friendliness and so on. These features are subjective by definition, and this, in the terms of science, makes them harder to evaluate.

The role of evaluation in HCI probably does not need explanations. In fact, in this field it may play even a bigger role than in others, as here investigating human's reactions and opinions towards computers is the main research content. In simple words – even if we construct a very complex, sophisticated system, it will mean literally nothing if the users will not want to interact with it, and evaluation is the only way to actually check it.

In our research we focus on the conversational side of HCI. In this paper we first investigate the field of non-task oriented conversational systems (so-called "chatterbots") evaluation methodology. As this topic is generally quite neglected, we propose our own methodology, which was tested in experiments we conducted during our research.

One evaluation method proposed in this paper (see **7.3**) is automatic and bases on emotion recognition from the textual layer of speech. To perform such analysis, we used an emotiveness analysis system for Japanese, in order to detect users' emotions towards conversational systems, by analyzing chatlogs from the interactions. Thus, it can be stated that the emotiveness analysis system is a sort of an AI tool, automatically performing evaluation experiments of conversational systems.

In the following sections we describe the background of our research and the methodological gap in the field of chatterbots evaluation. Next, we introduce our methodology (created on the basis of experience from our research so far), describe all used methods, discuss their usability and drawbacks. We also present some methods that we have not used yet, but that seem applicable in this methodology, such as brain scanning (detecting users' reactions directly from their brains). Finally, we point out several issues that need to be addressed in the future.

The methodology presented in this paper was used in our research to evaluate conversational systems. However, we believe that – with some minor changes – it can be more widely applied in evaluation of other aspects of HCI.

2. Conversational Systems

In this section we briefly review the state of the art in the field of conversational systems,

with particular regards to AI tools commonly used in research on such systems (see **2.2**).

2.1. *Two types of conversational systems*

There are two engines[†] commonly known types of conversational systems: task- and non-task oriented. Systems belonging to the former type are usually designed to conduct the conversation aiming at achieving a particular, well-defined goal. Good examples of such systems are information kiosks or virtual tour guiding agents.

Systems of the latter type, on the contrary, do not have one defined goal. They are commonly known as "chatterbots" - and, in fact, the name perfectly describes what such systems do: they chat with users. This is why they are called "non-task-oriented" systems – however, this diversion seems to us quite arguable, for here we face a question: does talking for pleasure or for amusement really mean the lack of goal in the conversation? If we assume it to be true, we actually question the sense of creating such systems – why make something that has no purpose? Meanwhile, although chatterbots are not designed to complete specified tasks (at least – in such understanding as information agents etc.), they are an important issue in nowadays computer science and thus should become a subject of more intensified research, including evaluation methodology.

2.2. *AI Tools in Conversational Systems*

In this section we provide a brief literature review on conversational systems, with particular regard to most common AI tools used to improve their performance. Table 1 presents a summary of this review, including tools used in systems described in this paper: Modalin and Pundalin (two chatterbots) and ML-Ask (emotiveness detector). The latter itself can be also considered an AI tool, as it was used to automatically evaluate the two chatterbots' performance (see **7.3**).

As the focus in conversational systems is obviously laid on NLP, it can be stated that most of them use techniques like "natural language understanding" or "natural language generation". These might seem overly generalized, but we believe they generally describe groups of AI techniques and tools used to understand and produce texts during interaction with users.

Table 1 AI tools used in conversational systems – summary (on selected examples)

Conversational system	AI tools used
Eliza (1966) [2]	natural language understanding, natural language generation, pattern matching, heuristics
SHRDLU (1972) [3]	natural language understanding, natural

[†] Some works mention three types of systems, placing virtual humans between chatterbots and task-oriented systems (e.g. [1])

	language generation, pattern matching, parsing, knowledge management
COMET (1991) [4]	natural language understanding, natural language generation, graphics generation, parsing
TRAINS (1995) [5]	natural language understanding, natural language generation, parsing, pattern matching, decision trees
Circuit-fix-it (1997) [6]	natural language understanding, natural language generation, pattern matching, parsing, speech recognition, speech synthesis, schemata
SmartKom (2003) [7]	natural language understanding, natural language generation, speech recognition, speech synthesis, schemata, pattern matching, parsing, gesture recognition, face recognition, function modeling, discourse modeling, context modeling, lexicon management
UAH (2005) [8]	natural language understanding, natural language generation, parsing, speech recognition, speech generation, dialogue management, database management, pattern matching, context processing
Let's Go system (2005) [9]	natural language understanding, natural language generation, parsing, speech recognition, speech synthesis, dialogue management, pattern matching
Kim et al. (2006) [10]	natural language understanding, natural language generation, semantic Bayesian network, information retrieval, pattern matching, parsing, dialogue management, knowledge management
ISU (2006) [11]	natural language understanding, natural language generation, reinforcement learning, parsing, dialogue management, pattern matching, speech recognition, speech synthesis, N-gram models

Our systems - Modalin, Pundalin (2008-2010)	natural language understanding, natural language generation, parsing, dialogue management, pattern matching, information retrieval, web mining
Our system - ML-Ask emotiveness detector (2008-2010)	natural language understanding, database queries, pattern matching, affect analysis

One of the first and best known dialogue systems is Weizenbaum's Eliza [2], a quasi-psychoanalyst chatterbot, which interacts with users via text. After the user types in an utterance, the system processes it and generates a response, using simple heuristics. It conducts conversations by asking questions rather than directly responding to users' utterances. Eliza uses a simple pattern-matching mechanism to understand texts inputted by users – if, for instance, user's utterance contains the word "mother", the system recognizes it as a sentence about a family member, and next generates a inquiring response including that particular word, like "Tell me more about your mother".

Despite being quite simple, Eliza itself was an interesting venture, also due to the fact that it was non-domain-restricted. It was said to deceive some users and make them believe they are talking to a human interlocutor, however, in the long run, dialogue with a system that mostly asks questions was said to be rather boring.

There are, however, systems that operate on higher level, within restricted domains. An ancestor of such domain-oriented systems, albeit not a very sophisticated one, is Winograd's SHRDLU [3], which simulated actions of a robot interacting within a "world of blocks", i.e. a domain containing different colored and shaped blocks, which could be placed on a table or put in a box. SHRDLU could perform conversations about the blocks world with users, and used sentence parsing of input commands and questions to understand what users' utterances. Although its replies were fairly limited or even boring, it still could answer correctly about the status of the blocks world, showing what could be called limited understanding of its virtual environment.

Another example of a domain-oriented dialogue system is COMET (COordinated Multimedia Explanation Testbed), developed by Feiner [4], which generated textual and graphical explanations for maintenance and repair of a certain type of radio equipment. This work is important in that it represents what is called a multimodal type of dialogue systems, generating output on more than one interaction level (here – text and graphics).

Also worth mentioning is Allen et al.'s TRAIN system [5], cooperating with users to schedule trains. It consists of three components: input module, responsible for processing user utterances, planning module, responsible for the train schedule composition, and output module, which generates responses, basing on the outcome of the planning module. While employing the usual set of AI tools (natural language understanding and generation, parsing, pattern matching, decision trees), this system is interesting due to the

fact that it cooperates with the user in order to achieve a mutual goal, being more of a virtual assistant than a conversation partner.

Another dialogue system that would guide its users in accomplishing particular tasks is Smith and Gordon's Circuit-Fix-It Shop [6], assisting humans in fixing certain types of radio circuits. Using a set of predefined schemata, the system processed user input in order to extract such information as circuit type and problems that particular user is experiencing, and next tries to define what caused them. It can detect errors caused by dead batteries, missing wires etc. The interactions with the system were speech-based, as it employed both speech recognition and speech synthesis. Using robust parsing, able to process also grammatically incorrect utterances, the system worked quite well, although its vocabulary was restricted to only 125 words.

The COMET system, mentioned above, represented what was called a "multimodal" type of dialogue systems, interacting on textual and graphical level. Some works go even further, such as the SmartKom system [7] – a personalized interface agent using speech and natural gestures in interactions with users. Apart from other "classical" AI tools used in most other systems, it also employed gesture recognition and face recognition, thus bringing the interaction to a visual level. Another important features of this system are sophisticated discourse modeling (including discourse memory), context modeling and well designed lexicon management.

Among other works in this field, we can mention also the UAH University on the Line, which provides spoken access to academic information at Universidad Al Habla [8]. It employed most of popular in such system AI tools, such as speech recognition / synthesis or pattern matching; it also employed quite a sophisticated context management procedure, capable of adapting the system's responses to the context, which makes the dialogue more natural. For instance, help messages generated by the system take into account current topic of the conversation.

The UAH University on the Line system was important in that it was available publically and tested as a real-life application. Another example of a widely accessible dialogue system was Raux et al.'s Let's Go system [9], which provides bus schedule information to the Pittsburgh population. Using also techniques like speech recognition / synthesis, the system was capable of performing successful real life dialogue with users in order to guide them in scheduling bus trips in Pittsburgh. The experiments showed that the system is robust and, after some improvements, can be used as a product.

Most systems mentioned above used more or less similar AI tools, while the system proposed by Kim et al. [10] represents a slightly different approach. It uses Semantic Bayesian Network (SeBN) to infer users' intentions during conversations. Using the probabilistic interference and the semantic interference, the system can understand and thus properly react to intentions displayed by users in their utterances. It was implemented into an information retrieval service for websites and verified by user-oriented evaluation experiments, which showed that most users were satisfied with the innovative SeBN-based method.

Another type of approach is presented by Lemon et al. [11], who proposed what they call an "ISU (Information State Update) Dialogue System", which uses reinforcement learning of dialogue strategies, and also has a fragmentary clarification feature. Primarily constructed in order to be able to collect data for Reinforcement Learning (RL) approaches to multimodal dialogue management, the ISU system can perform conversations within the in-car scenarios, focusing on providing the users with touristic information (e.g. hotels, bars, restaurants). The system can learn dialogue policies using the reinforcement learning technique, and this procedure can be called anytime, whenever the system needs to decide on its next dialogue move.

Also the systems presented in this paper use some AI tools in order to process user input and generate appropriate output. As their outlines are given in Section 6, here we only mention the tools that are used in our research. Both our chatterbots – Modalin (non-humor equipped) and Pundalin (humor-equipped) can understand and generate natural language. They both use MeCab – a sentence parser for Japanese [12], in order to acquire linguistic knowledge needed to fill in grammatical templates and generate responses relevant to user utterances. Both Pundalin and Modalin use the Internet as a source of lexical knowledge, and use it as a large, up-to-date data base.

Another system used in this work is ML-Ask Emotiveness Analysis System, which uses such techniques as database queries, natural language understanding or affect analysis in order to recognize users' emotions conveyed towards our chatterbots. To our knowledge, this is the first work in which automatic emotiveness analysis tool is used in conversational system evaluation. ML-Ask itself can be also considered an AI tool, as it was used to perform automatic evaluation of our conversational systems.

Thus, as summarized in this section, AI tools are commonly used in existing research on dialogue systems. Some of them might not be extremely sophisticated, but they seem to work quite well, in both non-domain and domain-oriented systems. Needless to say, there is still much to be done in this field and presumably some more sophisticated AI tools could be used to improve the dialogue quality even further.

3. Two Areas of Evaluation

Evaluation of dialogue systems depends strongly on their purposes and design. Most existing research projects (as, for instance, [6]) focus on areas which can generally be divided into two groups (see also Figure 1):

1) linguistic skills and/or technical quality focused;
2) non-linguistic skills focused.

The first area is basically common for both task- and non-task oriented systems. It concerns system's linguistic skills, such as grammar correctness, semantic naturalness or vocabulary richness, as well as the technical quality of interaction (time of system's reaction, voice recognition and generation, visual quality etc.). Such features are not very difficult to check. This area of evaluation is relatively objective.

The second area of evaluation differs for task- and non-task oriented systems. Those belonging to the former type are designed to achieve specified goals, and this, in most cases, can be seen as the priority in the non-linguistic skills focused aspect of evaluation. If we interact with a tourist information guide's, for instance, it is whether we were informed in the right way (i.e., the goal of the conversation was accomplished) that decides of our overall evaluation of the system's performance (in the non-linguistic layer). In our opinion, this presence of a specified goal, mutual for user and computer, makes evaluation in this area somehow easier to conduct, as criterion can be easily and relatively objectively verified.

In the case of chatterbots, however, non-linguistic skills focused area of evaluation cannot be as easily defined, as there is no mutual goal of the conversation, and it is the pleasure of having the interaction that counts in the first place. In other words, evaluation of such systems must focus on the user's impressions about these features of the interaction that make it more pleasant, natural and generally "better" in the eyes of users. Thus, by definition, such assessment has to be subjective.

However, this subjectivity does not necessarily have to be a drawback of chatterbot evaluation. In the end, this is what we want to check – the user's subjective opinion on the product. Another question is how we check it, or – which exactly features of interaction are worth investigating in order to give us desired results.

4. Methodological Gap

As mentioned above, non-linguistic skills focused area of chatting systems evaluation is by definition not objective. In fact, there are even no established quantitative metrics in this field, which leaves the researchers alone with their invention. In their work on evaluation of spoken language dialogue systems, Dybkjær and Bernsen [13] review numerous existing methods used to evaluate task-oriented systems. As for the non-task-oriented systems, they only state that "some of the usability issues (...) will clearly become irrelevant, such as sufficiency of task coverage, and others may suffer the same fate, such as informativeness. Instead, other issues may move into focus, such as conversational naturalness, turn-taking adequacy, and others" [13]. However, to our knowledge, no robust evaluation methodology that would normalize these issues exists, and all researchers are left with their own creativity and intuition.

Therefore, this paper can be seen as a contribution to this field. This is also the reason why we do not directly compare our methods with other existing research – there is simply no robust methodology we could compare to (we do, however, discuss some particular methods – see below).

In this paper we focus on non-linguistic area of chatterbots evaluation. However, the methods described here can also be applied in experiments on task-oriented systems, with some slight changes in their content. Although the priority in task-oriented system is the accomplishment of the task, it does not mean that such features as naturalness or human-likeness (which are of prior importance in the case of non-task-oriented systems) do not have to be taken into consideration. Also, these three types of methods can be used when

investigating task achievement degree and user's satisfaction – it can be done by users (first person oriented), by non-users (third-person oriented) and, to an extent, by using automatic evaluation systems.

5. Our Methodology

As briefly summarized in section **4**, the field of chatterbot evaluations is quite neglected, especially when the non-linguistic features are concerned. Thus, basing on our research, we decided to bridge that gap and propose our own evaluation methodology (see Figure 1).

5.1. *Non-task Oriented Systems*

In our research we focus on evaluations of the conversations between humans and non-task-oriented conversational systems. The easiest and most obvious method to do that is to ask users directly what they think about the interaction. After all, it is users who will interact with the system in the first place, and their direct opinion is of the highest importance. Therefore, asking the users questions about the interaction should be a compulsory element of all chatterbots (and HCI) evaluation experiments.

The well-known "Turing Test" [14] is also a sort of user-focused evaluation. Although its drawbacks are still widely discussed (summarization of these discussions can be found in many overview papers, e.g. Saygin et al. [15], it can be agreed that it allows to check one of non-linguistic features of interaction, which is "human-likeness". Antagonists of this method often claim that it focuses rather on deceiving users than on actual evaluation – however, we think that, to an extent, it can be used to check if the system's performance resembles human.

Although of high importance, evaluation conducted by users has also its drawbacks. First, even when conducted immediately after the interaction, it requires the user to remember his/her impressions from when the conversation took place. On the other hand, the alternative here would be to ask the user to evaluate the system during the conversation, which might distract him/her and negatively influence the smoothness of interaction. Another problematic issue is that we cannot be sure if the user is actually aware of his/her own feelings and impressions. Also, people are sometimes reluctant when telling about their feelings towards products, even during evaluation experiments. Therefore, although user-oriented evaluations remain of primary importance in HCI, conducting some complementary experiments to verify their outcome seems to be worth trying

Thus, in our research we employed two complementary, non-user focused evaluation methods: third person focused evaluation and automatic (emotiveness analysis based) evaluation.

In the former, system's performance is assessed by third person (non-user) participants. The easiest way to do it is to ask them to read and evaluate the chat logs by answering several questions on the interaction. If the evaluators do not know that one of dialogue participants was not human, they can evaluate both of them on equal rights. This,

in turns, opens a new possibility of comparative evaluation, in which scores of human speakers and systems can be compared (the smaller the differences between users and systems, the higher level of human-likeness).

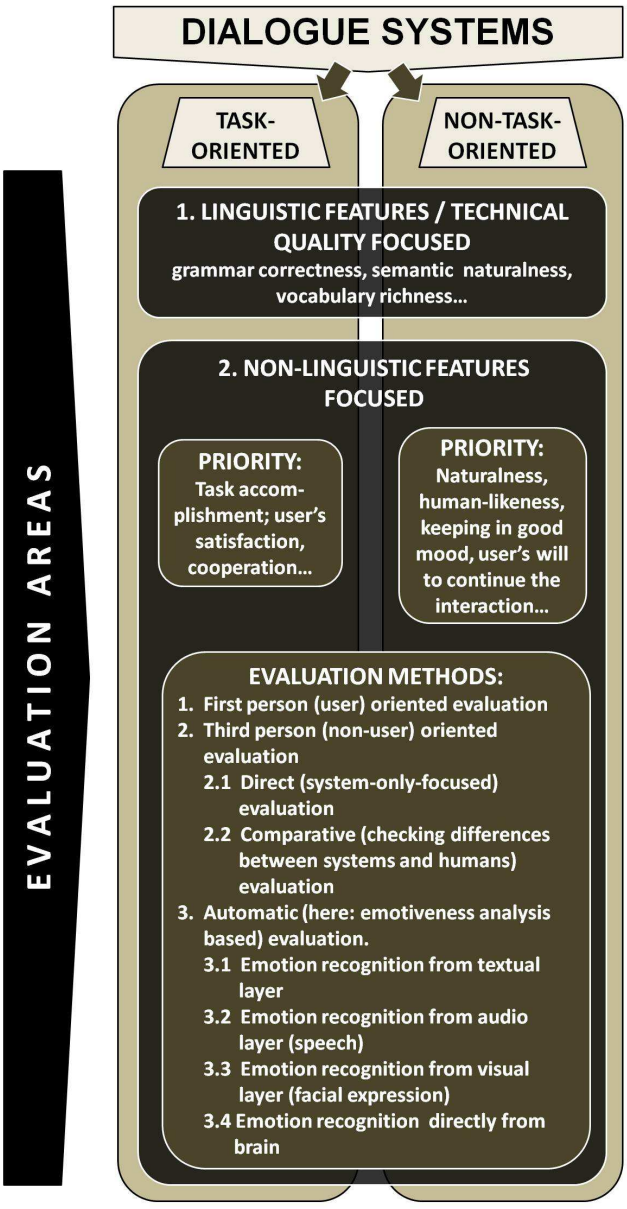

Fig. 1. Evaluation areas of task- and non-task oriented dialogue systems – proposal of methodology

In the latter (automatic emotiveness analysis based) method, users' utterances (reactions to system's output) are analyzed with emotiveness analysis system (ML-Ask – see **6.3**) in order to check changes in users' emotional states and general attitude towards the system (based on the "affect-as-information" approach – see **6.3**).

5.2. *Task-oriented Systems*

As mentioned above, in our research we focus on chatterbots. However, as showed on Figure 1, the evaluation methods in the non-linguistic features focused area should be the same for both task- and non-task oriented systems. Needless to say, as the priorities are different, the contents of evaluations should also differ – nevertheless, the main methods remain more or less the same. Emotiveness analysis, for instance, can be used to measure users' satisfaction level after accomplishing (or not) a task. First- and third-person oriented methods would also prove useful here.

5.3. *Other Methods*

The methods mentioned above have been tested in our experiments (see below). However, we have also some ideas that are still to be used. The most important one is a method based on brain scanning, in which users emotions during the interaction with the computer would be detected directly from their brains (see **7.5** for details).

Obviously, users' emotional states can be detected also from other layers of behavior. In our research we apply only textual layer analysis (as our chatterbots are text-based, this seems to be the most proper method here) – however, there are also research on recognizing emotions from speech or facial expressions. Although these methods (briefly summarized in 7**.4**) base on completely different features, ways of applying them to the emotiveness analysis-based evaluation are generally the same – we can derive users' attitudes to the system by analyzing their emotional states.

5.4. *Methodology - summary*

To summarize the methods listed above, we have:

1) **First person (user) oriented evaluation**
2) **Third person (non-user) oriented evaluation**
   - 2.1) **Direct (system-only-focused) evaluation**
   - 2.2) **Comparative (checking differences between systems and humans) evaluation**
3) **Automatic (here: emotiveness analysis based) evaluation.**
   - 3.1) **Emotiveness recognition from textual layer**
   - 3.2) **Emotiveness recognition from audio layer**
   - 3.3) **Emotiveness recognition from visual layer (facial expressions)**
   - 3.4) **Emotiveness recognition directly from the brain**

Below we describe and discuss the methods from non-linguistic area, using examples from our earlier research.

6. Systems Used in this Research

Most of the methods of chatterbots' evaluation presented in this paper have been used in our research on humor-equipped talking systems. In this section we briefly describe our three chatterbots (one baseline systems – Modalin - and one humor equipped system Pundalin), along with the Emotive Elements/Emotive Expressions Analysis System (ML-Ask), used in the automatic evaluation experiments (see **7.3**).

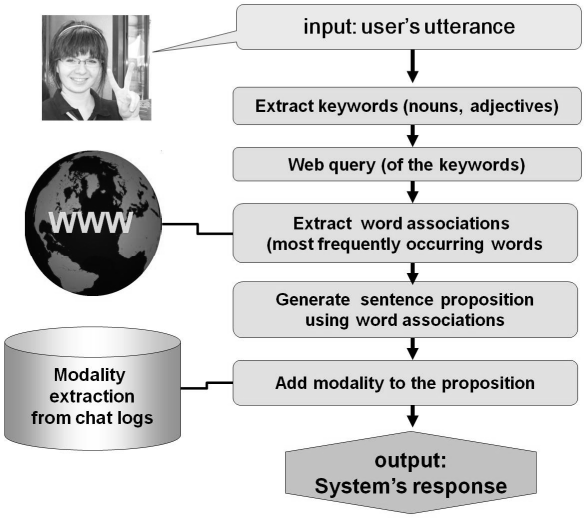

Fig. 2. Baseline chatterbot (Modalin) – algorithm outline

6.1. *Baseline Chatterbot Modalin*

The baseline system in our research is a Japanese, text based chatterbot called "Modalin" (developed by Higuchi et al. [16]), which uses the Internet to extract word associations for user utterances and adds modality to the generated responses. For example, from the user input phrase:

"- *Nanika sukina tabemono aru*? (What food do you like?)",

the system extracts a keyword: "*tabemono*" ("food"), checks its associations in the Internet (using snippets – small descriptions of pages, found in query engines), finds the word "*oishii*" ("tasting good") as the closest one to the keyword (one with the highest co-occurrence rate), and uses it to generate a sentence. Then, it adds modality (*"maa"* – "well") to the generated phrase, and outputs the response, which in this case is:

"- *Maa, tabemono-wa oishii desu*. (Well, food tastes good.)."

The correctness of generated sentence is once again checked in the internet.

In the evaluation experiment (linguistic skills focused – see [16]), human subjects assessed above 80% of the extracted associations as correct. The experiment also showed that adding modality to the system's response greatly improved the performance.
The system's algorithm outline is presented on Figure 2.

6.2. *Humor-equipped Chatterbot - Pundalin*

The two chatterbots described above were used to create a joking (pun-telling) chatterbot for Japanese, named "Pundalin".

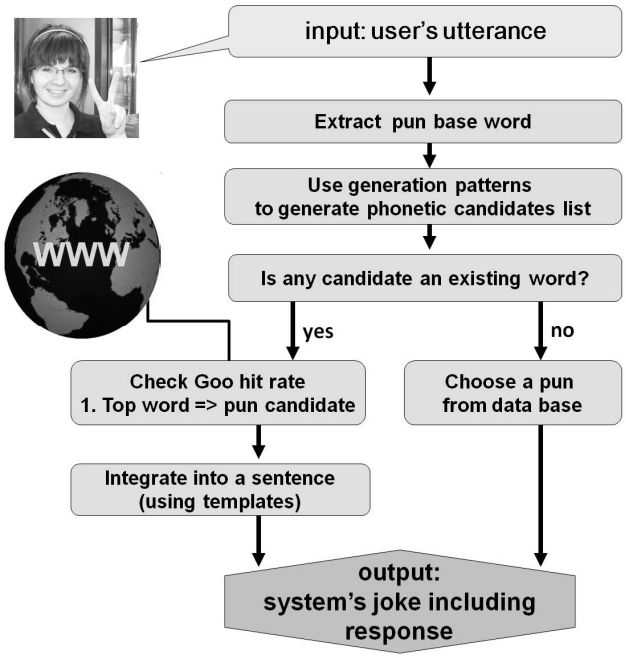

Fig. 3. Humor generating algorithm (PUNDA) – algorithm outline

The pun generating engine PUNDA used in this research was developed by Dybala et al. [17]. It also uses the Internet. From a user utterance it extracts a base word and transforms it using Japanese pun phonetic generation patterns to create a phonetic candidate list. Then, it checks all candidates using the Goo search engine[‡], and chooses the one with the highest hit rate. Next, it uses manually created sentence templates (such as "speaking of A, it's B) to generate a response.

Below is an example of the system working:

[‡] www.goo.ne.jp

User: *Kaeru daikirai!* ("I hate frogs!")
Base word: *kaeru* (a frog)
Pun candidate: *tsukaeru* (to be able to use)
Template: *A to ieba B ne* (Speaking of A, it's B, right?)
System: *Kaeru to ieba tsukaeru yo ne!* („Speaking of frogs, they can be useful, can't they?")

If no candidate was found for the user's input, the system randomly selects a pun from our pun database.
The system's algorithm outline is presented on Figure 3.

The algorithm described above was combined with the chatterbot (Modalin) to create a talking system which tells jokes, named "Pundalin". As far as joke timing is concerned, at first we decided to apply a very simple rule – in every third turn of the conversation, the first talking system's output was replaced by a joke-including sentence, generated by the joking system. In other words, the user's every third utterance becomes input for the joke generator, which generates an appropriate pun for it.

In a later version of the system, the "every-third-turn" rule was replaced by a decision-making procedure, performed by the ML-Ask Emotiveness Analysis System (see **6.3**). The system detects users' emotional states, and if they are negative – it decides to tell a joke in order to try making them feel better. This work is currently in evaluation stage.

6.3. *Emotiveness Analysis System – ML-Ask*

In our research we used ML-Ask Emotive Elements/Emotive Expressions Analysis System for Japanese (developed as a part of one of our other research project [18]) to perform automatic analysis of chat logs acquired in the user-focused experiment (see **7.3**). Based on Ptaszynski's idea of binary classification of realizations of emotions in language [19], the ML-Ask system performs utterance analysis in two general steps:

1. **Determining general emotiveness (emotive/non-emotive), and**
2. **Specifying types of emotions found (in emotive utterances only).**

The system's algorithm outline is presented on Figure 4.

In the first step, if many of the evaluator's utterances were determined as emotive, it was assumed that he or she was emotionally involved in the dialogue. In Japanese, emotional engagement in the conversation suggests a tendency to familiarize with the partner – which, in this case, is the conversational agent.

In the second step, analysis of the specific emotions showed by the evaluators during the conversation provided us with the information of their feelings towards the system. If emotions detected by ML-Ask were positive or changing from negative through neutral (non-emotive) to positive during the whole conversation, the general sentiment towards the system was considered to be positive. If the emotions detected by ML-Ask were

negative or changing from positive through neutral to negative during the whole conversation, the general sentiment towards the system was considered to be negative [20].

In our research, we approach to emotion types and emotion expressions using two general dimensions: positive/negative and activated/deactivated [21]. Each type of emotions can be described using these two dimensions.

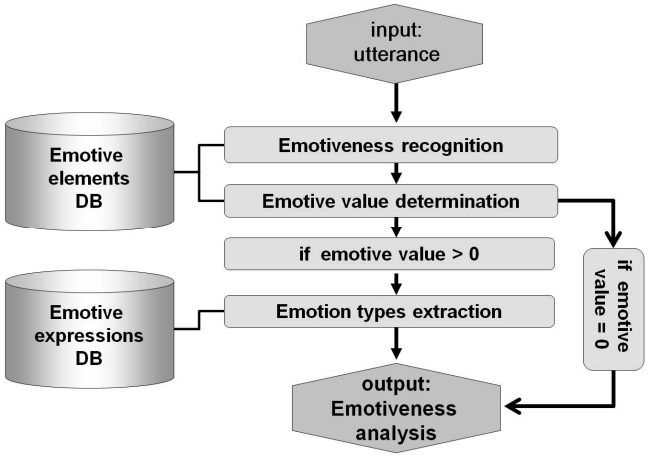

Fig. 4. ML-Ask System – algorithm outline

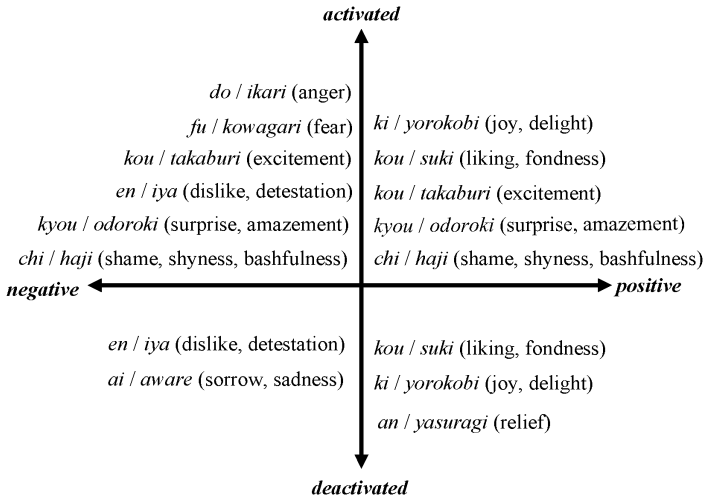

Fig. 5. Grouping Nakamura's classification of emotions on Russell's two-dimensional space [8]

The types of emotions in our research are based on Nakamura's Japanese emotion classifications (10 types) [22]. The proposed emotion types were projected on Russel's 2-dimensional model of affect [21]. The effect of this projection can be seen in Figure 5. Some types of emotions appear more than once, as they contain expressions from more than one quarter – for example, group "*kou/suki*" (liking, fondness) may contain both activated and deactivated expressions.

Affect-as-information

The idea of using emotiveness analysis to evaluate products is not new. Perceiving affect as information was proposed by Schwarz and Clore [23]. The main idea of this approach is based on a claim that humans use affect in the same way as any other criterion, namely by using the informational value of their affective reactions to form their opinions and judgments. This leads to an assumption that information about someone's attitude to a product can be derived from information about changes in his or her affective states during its usage. Basing on this assumption, we used the ML-Ask system to perform emotiveness analysis based evaluation of our systems (see **7.3**).

7. Evaluation Methods

This may sound completely obvious, but we need to state one thing before we start explaining and discussing our evaluation methods: the shape of evaluation experiment strongly depends of what we want to check. In the evaluation experiment we explore both linguistic- and non-linguistic area – however, we focus on the latter, as it is the role of humor in the interaction we want to check in the first place. For the same reason, some of the questions in our evaluation were directly related to humor and perceived funniness of the systems' utterances.

Taken these two aspects into consideration, in our evaluation we decided to explore such non-linguistic aspects of interaction as: human-likeness, the will to continue the dialogue, engagement in the conversation, funniness and emotive response. We also evaluated the linguistic area of systems' performance – however, in our research these results were of secondary importance.

These issues are described in details in below sections.

7.1. *First Person Oriented Evaluation*

As it is users who are the "clients" of our product, in the research on dialogue systems the first person oriented evaluation is to us of highest importance. Although not perfect, methods described in this section allow us to investigate the users' impression of the interaction with the system, in the possibly most direct way. As stated above, this evaluation is by definition subjective, but we do not see it as its drawback. Instead, we propose to accept the lack of objectivity as a natural fact in impression-relying evaluation. Individual differences are inevitable here, and their presence can be a trigger to construct more sophisticated, user-adapting systems. An idea of such system for humor-oriented chatterbots is described in one of our works [24].

There are two major methods of first person oriented evaluation: one in which user's impressions about the interaction are checked by filling out a questionnaire or conducting an interview, and another one conducted using the Turing Test. The latter for years has been the subject of many discussions and actually received a lot of criticism. Therefore, as this method is probably the best known of all mentioned in this paper, we are not going to discuss it in details here. We would only like to mention that – comparing to the questionnaire/interview method – Turing Test does not give us detailed and measurable information about the interaction, while its only aim is to check if the system is sophisticated enough to deceive users about its not being human.

Therefore, in our research we used the former first person evaluation method to conduct the first person oriented evaluation experiment. Human evaluators were asked to perform a 10-turn dialogue with Modalin (non-humor-equipped), and Pundalin (humor-equipped system). No topic restrictions were made, so that the talk could be as free and human-like as possible. Thus, the variety of utterances was quite big – most of them, however, resembled normal (human-like) beginning of conversation, for example: "What did you do yesterday?", "May I ask you a question?" or "It's hot today, isn't it?"

There were 13 participants in the experiment, 11 male and 2 female; all of them were university undergraduate students. After talking with both systems, they were asked to fill out a questionnaire about each agent's performance. The questions concerned both linguistic (B-D) and non-linguistic (A, E-H) areas of interaction. The questions were:

A) Do you want to continue the dialogue with the system?
B) Was the system's talk grammatically natural?
C) Was the system's talk semantically natural?
D) Was the system's vocabulary rich?
E) Did you get an impression that the system possesses any knowledge?
F) Did you get an impression that the system was human-like?
G) Do you think the system tried to make the dialogue more funny and interesting?
H) Did you find system's talk interesting and funny?

The replies to the questions were given on 5-point scales with some explanations added. Each evaluator filled out two such questionnaires, one for each system. The final, summarizing question was "Which system do you think was better?"

Generally, the questions can be divided into two groups: 1) concerning the linguistic qualities of the conversation (B-D); and 2) concerning the general impression of the interaction and the systems' performances, in this case focusing on the role of humor. The last question ("which system do you think was better?"), although quite general, seems to be a natural and intuitive way of comparing two or more similar entities. In our research we want to create a system that will interact better with humans – and this makes this question even more relevant.

Statistical significance of the results was calculated using the Two Paired Sample Wilcoxon Signed Rank Test (as the data was paired, but did not have a normal distribution). The results are summarized in Table 2.

Table 2 User's evaluation – results for Modalin and Pundalin for detailed questions (see **7.1**). Answers were given in a 5-point scale.

Question	Modalin	Pundalin	Difference	P value
A	2.62	3.38	**0.76**	**<0.05**
B	2.15	2.92	**0.77**	**<0.05**
C	1.85	2.69	**0.84**	**<0.05**
D	2.08	3.00	**0.92**	**<0.05**
E	2.15	2.85	**0.70**	**<0.05**
F	2.38	3.31	**0.93**	**<0.05**
G	1.92	4.15	**2.23**	**<0.05**
H	2.46	4.08	**1.62**	**<0.05**
Which better?	15%	85%		

The results show that the system with humor received higher scores in both linguistic and non-linguistic areas. As for the former, it may seem unusual that the presence of humor improved the system's linguistic skills – this fact, however, could have been caused by the fact that Pundalin uses fragments of human created sentences and jokes from the data base, which obviously are more correct than those automatically generated by the computer.

Also in the non-linguistic area all results point at the humor-equipped system. Users wanted to continue the conversation to a higher degree with Pundalin than with Modalin, perceived Pundalin as more human-like, knowledge-possessing, funny and generally better than Modalin [17].

Results for all questions were found significant on 5% level.

Discussion

As mentioned above, the first person oriented method to be the best and most direct way of evaluating the system. Such a method (questionnaire/interview) was also used by Bernsen and Dybkjær in their experiments on NICE - a system for spoken and gesture interaction with life-like fairytale author Hans Christian Andersen [25]. The content of questions was slightly different from those we used, as the embodiment of the system and the usage of gestures also needed to be addressed. However, the general tendency in NICE's evaluation was consistent with ours: most questions concerned users' subjective impressions about the interaction (such as: "How did it feel to talk to the system?", "What was bad about your interaction with the system?" or "What was good about your interaction with the system?" [25]). The biggest difference between our and Bernsen and Dybkjær's evaluation was qualitative – e. g., the answers in their experiment were given freely by the users, without any quantitative scale. Users' responses were manually analyzed and their descriptive summarization is the result of the experiment. Such non-quantitative methods surely can give us a deeper insight into the users' impressions about

the system – however, it is quite hard to use when comparing with other systems. Therefore, we think that the qualitative methods should be used rather in preliminary experiments, in order to receive feedback from the users, than to evaluate the final product.

While of high importance, user-focused evaluation with the use of questionnaires or interviews also has its drawbacks. First, even when conducted immediately after the interaction, it requires the user to remember his/her impressions from when the conversation took place. On the other hand, the alternative here would be asking the user to evaluate the system during the conversation, which might distract him/her and negatively influence the smoothness of interaction. One way to solve this problem is to conduct also the third person oriented experiment (see **7.2**), in which chat logs from first person experiment are evaluated by non-user participants. If the chat logs are printed, the whole conversation can be referred to during the evaluation, and therefore it does not require the participants to remember what was said and what they felt about that. This approach, however, has also several drawbacks, which are discussed in section **7.2**.

Another problematic issue of the first person oriented evaluation method is that we cannot be sure if the user is actually aware of his/her own feelings and emotions. Obviously, they have influence on the user's impression towards the system – however, if we were to check his/her specific feelings, asking direct questions may not give us the exact answer, as sometimes it is not easy to realize or talk about one's own emotions.

In our research we partially solved this problem by using the automatic emotiveness analysis based evaluation. Even if the users do not fully realize their feelings and do not reflect them in the evaluation, there is a chance that they will be reflected in their words, voice or facial expressions. Therefore, emotiveness analysis of the users' utterances can be seen as a good complementary evaluation method.

7.2. *Third-Person Oriented Evaluation*

In the first person oriented evaluation we ask the users directly about their impressions about the system. However, as mentioned above, one drawback of this method is that the evaluation has to be conducted after the conversation. To solve this problem and double-check the results of the first experiment, we conducted an additional experiment, in which third person (non-user) participants evaluated the chat logs from the users experiment. The questions asked were similar to those used in the user-focused experiment – we only made minor adjustments. First, the word "system" was changed to "dialogue" (in some cases - "Speaker"), as we did not want the evaluators to know that some of the utterances were generated by a computer system. In the chat logs given to the third person evaluators, dialogue participants were called "Speaker A" for the user and "Speaker B" for the system. In addition, question F (about human-likeness) was deleted, as it would also reveal that at least one speaker was not human. Also, in questions B, C, D, E, G and H we added two options: 1) "Speaker A" and 2) "Speaker B" – so that the dialogue participants would be evaluated separately. Thus, the list of questions used in this experiment goes as follows:

A) Do you want to read the continuation of the dialogue?
B) Was Speaker A/B's talk grammatically natural?
C) Was Speaker A/B's talk semantically natural?
D) Was Speaker A/B's talk vocabulary rich?
E) Did you get an impression that Speaker A/B possesses any knowledge?
F) <Deleted>;
G) Do you think the Speaker A/B tried to make the dialogue more funny and interesting?
H-1) Did you find the dialogue interesting and funny in general?
H-2) Did you find Speaker A/B's talk interesting and funny?

After completing the detailed questionnaire, the evaluators answered the final question, the same as in the previous experiment - "Which dialogue did you find most interesting and funny?" (we used the Japanese word *omoshiroi*, which can mean "interesting" or "funny", and is generally positive in meaning [22]).

The chat logs were divided into 13 sets. Each of them included one Modalin and one Pundalin dialogue. Each set was evaluated by 5 participants, which makes a total of 65 evaluators, all of which were university students. [17].

The results were analyzed using two methods mentioned above: direct and comparative.

7.2.1. *Direct*

In this method we only take into consideration the results for systems' (Speaker B's) utterances, and compare them for both humor- and non-humor equipped system, as we did in first person evaluation. The results of this method are summarized in Table 3.

Although the differences here were not that clear and significant as in the user-focused evaluation, the tendency is still visible. The humor-equipped system received higher scores in all categories. Only in two cases (D and F) were the differences found to be statistically significant – however, the results for the general question show that even if there is not much difference, the evaluators still chose dialogues with humor (69% vs. 31%) [17].

Discussion

As shown in the Table 3, the results are generally consistent with those of first person oriented experiment. Thus, it can be stated that the direct third person oriented method can be used to evaluate chatterbot's performance. However, low differences and insignificance in this experiment require more detailed discussion.

One of possible explanations of this phenomenon is that when users are talking with the systems, they are usually quite impressed by the very fact that a computer can talk. This very fact may positively influence the results. In this understanding, third person oriented evaluation seems more objective, since the evaluators were not the participants of the interaction, and thus had more distance to the subject of evaluation. Also the fact

that they did not know that one of the speakers was a computer system was probably not without meaning (this issue is discussed in **8.1.2**).

Albeit the relative "objectiveness" of the third person evaluation (more distance towards chat logs than towards conversation partner), this method has also several drawbacks. The major one is that, as mentioned above, it is the user that has to be satisfied in the first place. They will use the system and it is their opinion that counts the most. Of course, the more severe and diversified the evaluation, the more information about our systems and enhancements needed we can get; however, to evaluate the final product it is still the first person oriented evaluation that should be of primary importance. On the other hand, third person evaluators are also potential users, and if we are going to construct machines that would be our companions, able to interact freely with ordinary people (not only selected group of users), opinion of such potential users also should be taken into consideration. Apart from that, third person oriented methods can be used to double-check the results or to acquire feedback leading to the system's development.

Table 3 Third person evaluation – results for Modalin (non-humor equipped system) and Pundalin (humor-equipped system). Answers were given on a 5-point scale.

Question	Modalin	Pundalin	Difference	P value
A	2.60	2.89	**0.29**	>0.05
B	1.78	2.09	**0.31**	>0.05
C	1.48	1.69	**0.21**	>0.05
D	2.03	2.38	**0.35**	**<0.05**
E	1.87	2.13	**0.26**	>0.05
F	X	X	X	X
G	2.51	2.91	**0.40**	**<0.05**
H-1	2.88	3.19	**0.31**	>0.05
H-2	2.73	3.16	**0.43**	>0.05
Which better?	31%	69%		

7.2.2. *Comparative*

As mentioned above, in our third person evaluation experiment we referred to the speakers in the chat logs as Speaker A (the user) and Speaker B (the system). In the direct evaluation we took into consideration only the results for Speaker B, while in the comparative evaluation we calculated the differences between the systems and the users. Statistical significance of all scores was calculated. The results are summarized in Tables 4 and 5.

Table 4 Results for Modalin for detailed questions in third person evaluation (differences between users and systems). Question F was deleted (see **7.2**), and question A did not include separate options for Speakers A and B. Minus values mean that Speaker B (the system) received higher scores than the user.

Modalin				
Question	**User**	**Modalin**	**Difference**	**P value**
B	3.30	1.78	**1.52**	<0.05
C	2.94	1.48	**1.46**	<0.05
D	2.92	2.03	**0.89**	<0.05
E	3.13	1.87	**1.26**	<0.05
G	2.54	2.51	**0.03**	>0.05
H-2	2.85	2.73	**0.12**	>0.05

Table 5 Results for Pundalin for detailed questions in third person evaluation (differences between users and systems). Question F was deleted (see **7.2**), and question A did not include separate options for Speakers A and B. Minus values mean that Speaker B (the system) received higher scores than the user.

Pundalin				
Question	**User**	**Pundalin**	**Difference**	**P value**
B	3.18	2.09	**1.09**	<0.05
C	3.00	1.69	**1.31**	<0.05
D	2.81	2.38	**0.43**	<0.05
E	2.97	2.13	**0.84**	<0.05
G	2.52	2.91	**-0.39**	<0.05
H-2	3.09	3.16	**-0.07**	>0.05

As shown in above tables, the results show that the humor-equipped system differs less from humans than the non-humorous one. In other words, the difference between humans and Pundalin was smaller than the difference between humans and Modalin. In our research this is especially important for questions D-H, which belong to the non-linguistic area of evaluation. Looking at the results, we can see that the system with humor actually made more effort than humor to make the dialogue interesting. The fact that Pundalin surpassed the users in this category can be interpreted as not necessarily positive, as trying too hard may also be annoying. However, knowing that the system was assessed as generally better both by the users and third person evaluators, we can assume that the attempts to make the conversation more interesting were rather appreciated than disliked.

Discussion

From these results, a conclusion can be drawn that the system which differs less from humans can be seen as more human like. This assumption is consistent with the results of

the first person oriented experiment (see **7.1**, question F). Obviously, more research on the issue of human-likeness is needed – however, we think that the method suggested here is also an option to check how close to the human level the system is.

This method, albeit innovative, has also several drawbacks. The main one is the same as in case of the direct evaluation – it may be slightly less subjective, but it is also less direct, and does not involve the users. However, the consistency with the results of user-oriented experiment shows that it can be used as a complementary method of evaluation.

In previous sections we mentioned the Turing Test, as probably the best known (although arguable) method of checking the system's human likeness. The Turing test is a first person oriented method, in which the users have to tell if the interlocutor is a human or a computer. However, it should be possible to conduct its third person oriented version, in which the evaluators would read the chat logs and guess the identity of speakers. Obviously, third person oriented Turing Test would have the same drawback as other third person oriented methods – however, we believe that it can be a good complementary method and as such may be worth trying.

7.2.3. *Coping with the same output*

When conducting experiments in which two (or more) similar systems are compared to each other, it would seem reasonable to investigate how they cope with the same input. This, however, would require the users to perform interactions in exactly the same scenarios, i.e. to input always the same utterances, regardless to the systems' responses, which, needless to say, would be rather pointless. Thus, we decided to focus only on short parts of chatlogs, where, on a distance of 3 turns, differences between the two systems could be visible without loosing the sense of the dialogue. For each Modalin dialogue from the first-person oriented experiment, first three turns were automatically selected, and user's third utterance was used as an input for the PUNDA pun generator to generate a humorous response. As after altering the third system's response, the rest of the dialogue became irrelevant, for this experiment's purpose it had to be ended after this turn.

The evaluators were 65 university students, 37 male and 28 female. They were not told of the origin of dialogues and apparently did not know that some of evaluated utterances were generated by computer. Each participant evaluated 1 set, including 3 short dialogues: Modalin only, Modalin plus PUNDA (with the third system's response replaced by PUNDA's joke) and Pundalin. There were 13 sets of dialogues, and each of them was evaluated 5 times. For the dialogues were too short to be evaluated in details, in the questionnaire one general question was: "Which dialogue do you find most interesting and funny?".

The results of the experiment were generally consistent with those of other experiments. Among 65 evaluators, only 10 (15.4%) responded that Dialogue 1 (Modalin only) was most interesting and funny. 20 (30.8%) pointed out Dialogue 2 (Modalin plus PUNDA) and 35 (53.8%) – Dialogue 3 (Pundalin only). This means that each of humor containing dialogues received evaluation clearly higher than non-humor dialogue (84.6%

as a sum of both humorous dialogues) - see Figure 6. It may also suggest that the manner in which PUNDA deals with the same input is generally better and more liked by the users, as Dialogue 2 received twice as many votes as Dialogue 1. [17]

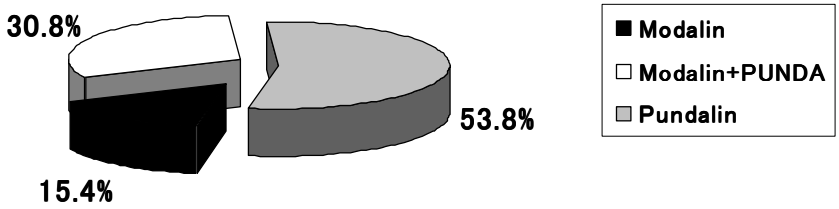

Fig. 6 Third person evaluation - results for the question "Which dialogue do you find most interesting and funny?" (short dialogues) [17]

Discussion

Although presenting only short parts of dialogues to evaluators might seem of little importance, it can provide us with valuable data on how both systems cope with the same input. Results point at the system with humor as performing better also in this category. As far as the method is concerned, it does not require any additional effort from users, and bases only on chatlogs acquired in the first-person oriented experiments. Thus, it seems a usable option to evaluate both systems' manner of dealing with the same user utterances.

7.3. *Automatic Evaluation – Emotions from Text Recognition*

In the above sections we discussed some drawbacks of first and third person oriented evaluation methods. In the former, completing a questionnaire has to be either delayed in time (after the interaction) or interrupt the dialogue (during the interaction). The latter method is less subjective, but indirect. In addition, both of them are time- and effort-consuming, and it is often necessary to reward the evaluators, either with money or with something else (e.g. credit or some extra points for students).

These problems can be solved by using automatic methods, in which the system's performance is evaluated by another system. In our research we used ML-Ask Emotive Elements/Emotive Expressions Analysis System to perform automatic analysis of the textual layer of the chat logs acquired in the user-focused experiment. The method does not completely exclude the presence of users – however, all they have to do is to chat with the systems, and there is no need to fill out any questionnaires or take part in interviews.

Using some simple AI techniques, such as information retrieval or pattern matching, ML-Ask system itself can be also considered a sort of AI tool, usable for user modeling, as it can be used to automatically recognize users' emotions towards HCI systems.

7.3.1. *General Emotiveness*

In the first step, general emotiveness (emotive/non-emotive) of users' utterances was analyzed by the ML-Ask system. As shown in Figure 7, most of the users showed more emotions towards Pundalin than towards Modalin, which means that they were generally more emotively involved in the conversation with the system which used humor [20]).

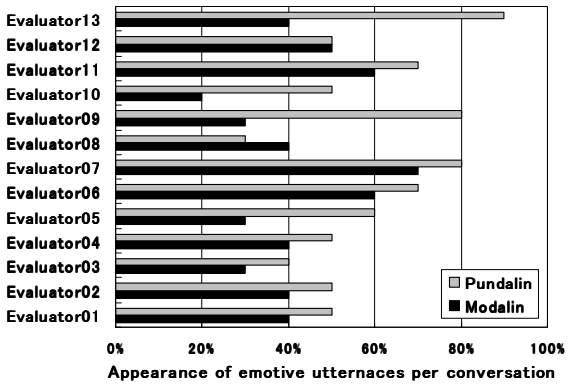

Fig. 7. Average percentage of appearance of emotively engaged utterances for all human evaluators in conversations with Modalin and Pundalin. [26]

7.3.2. *Valence and Specification*

In the next step of the evaluation, the chat logs were analyzed to check the specific types of emotions in the users' utterances. Figures 8 and 9 show the results projected on the Russel's two-dimensional space (see also Figure 5).

As shown above, while most of emotions towards Modalin were negative and activated (75%), for Pundalin the proportion was opposite (45% of positive and activated, 78% of positive emotions in total). In this experiment, no negative deactivated emotions were found neither in humor-equipped, nor non-humor- equipped system's chat logs.

7.3.3. *Positive / negative Engagement*

The correlation between speaker's emotiveness and conversation engagement was proved in various researches (e.g. [27], [28]). This knowledge was also used by Yu et al. [29] in their research, in which they measured engagement level basing on emotion recognized in user's speech. The approach is quite similar to ours – however, the efforts made by Yu et al. do not include explicit proofs for the correlation between engagement and emotiveness.

Thus, to investigate the correlation between engagement and emotiveness of conversation participants, we conducted a small-scale experiment [26], in which we proved that dialogues in which participants' engagement level was assessed as high, were also assessed as more emotive.

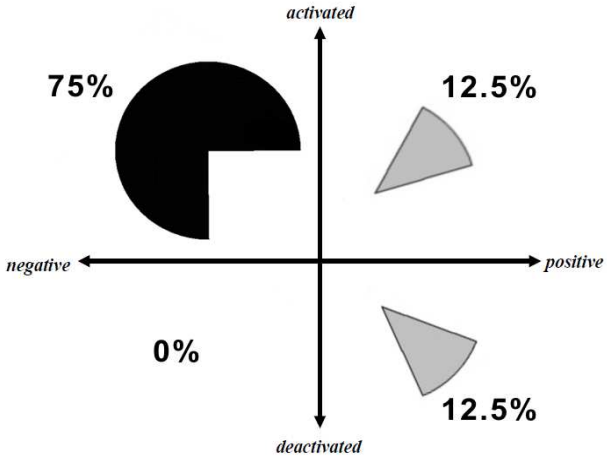

Fig. 8. Projection of emotive analysis of users' emotions types on Russell's two-dimensional space – Modalin (without humor) [26]

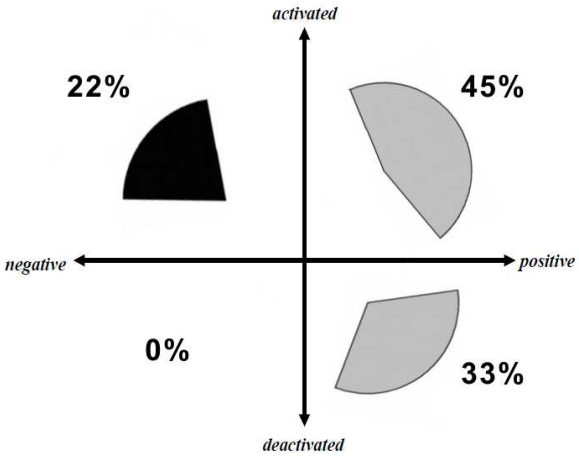

Fig. 9. Projection of emotive analysis of users' emotions types on Russell's two-dimensional space – Pundalin (with humor) [26]

In our evaluation experiment, conclusions about the users' engagement in the conversation were drawn basing on the results described in **7.3**. General emotiveness analysis results (see Figure 7) suggest that the users were more engaged in the conversation with Pundalin, as they showed more emotions towards it than towards Modalin. However, the engagement does not necessarily have to be positive, as the users might have reacted to the systems' utterances with negative arousal or irritation. This is

why we propose to distinguish between positive and negative engagement. To do that, in our research we base on emotive valence (positive/negative emotions – see Figures 8 and 9). The results show that the users were generally more positively and less negatively engaged in the conversation with Pundalin. This is consistent with the results of previous experiments, especially with the questions concerning the will to continue the interaction (see **7.1**), as this issue is strongly related to the engagement.

Therefore, it can be stated that automatic emotiveness analysis can also be used to investigate user's engagement and its types.

The role of activation/deactivation of emotions is still to be investigated.

Discussion

In this section we described automatic evaluation method used in our research. This method has a few significant advantages. First: it is automatic. This means that it does not require any additional engagement from the users in the experiment – they only have to perform a conversation with the system. We do not have to waste our and other people's time. The evaluation can be conducted at any time, not necessarily right after the conversation.

Second: as mentioned above, it is quite difficult to speak about one's own emotions and feelings, as we may not be fully aware of all of them. Not mentioning the fact that – if the evaluation is conducted after the conversation – users may not exactly remember what they felt some time ago. These feelings and emotions, however, can be revealed in the users' behavior during the interaction, which makes them possible to detect using recognition algorithms, from textual layer (as in our research) as well as from other layers (see **7.4**).

7.4. *Emotion Recognition from Audio and Visual Layer*

In our research we focus on recognizing emotions from the text. However, they can be also recognized from other layers, such as speech (voice) or facial expressions. In the below sections we briefly summarize these two areas of emotion detection.

7.4.1. *Audio Layer*

Most systems from the former group focus on prosodic features (e.g., pitch-related feature, energy-related features, speech rate) and spectral features (e.g., MFCC, cepstral features). Among them, pitch and energy seem to contribute most to affect recognition [30].

Some research projects in the field of emotion from speech recognition use data that are collected in an artificial way, namely by asking actors to speak prescribed utterances with certain emotions (e.g. [31]). Others turn to more natural data, such as phone call recordings (e.g. [32], [33]), meetings [34] or records of conversations with dialogue systems (e.g. [35], [36]). However, in such natural data, affect displays are often quite subtle and much more difficult to detect than in data prepared deliberately. Therefore, detection in this field is often limited to valence (positive / negative), without specifying

particular types of emotions ([33], [35], [36]). This is also one reason why the textual layer should be taken into consideration – as showed above, our system can also detect emotion types from naturally collected data. However, this does not mean that we are not planning to include the audio layer in our further research, as in the future our system will be converted from text-based to talking, with voice recognition and generation. Then, combining emotion recognition from text and audio (and, possibly, video) layers seems to be the best and most natural method to conduct the automatic type of evaluation.

7.4.2. *Visual Layer*

In the field of automatic facial affect recognition, most of existing works focus on Ekman's six basic emotions (joy, anger, fear, sadness, disgust and surprise) [37], due to their universality in our affective lives as well as to wider availability of test and training data. Also here one problem is that some researchers use data artificially created, often by actors (e.g. [38]). Such deliberately expressed emotions are often exaggerated, while natural expressions are much more subtle. Thus, many works recently turn to spontaneous facial displays – like recording of human-human conversations (e.g. [39], [40], [41]) or TV broadcasts [42]. There are also studies [43] focusing on distinguishing between posed and genuine emotions.

Most of existing works in this field are based on 2D spatio-temporal facial features, with two main subgroups: geometric features (like shapes and locations of particular facial points, such as eyes or mouth – e.g. [44]) and appearance features (facial texture changes, such as wrinkles or furrows - e.g. [45]). There are also a few works basing on 3D face models, focusing on such features as head movement [46]. The methods in this field still need improvement – however, the task is worth effort, as with the use of 3D models, the subject can be recorded in more natural conditions (i.e. does not have to sit still right in front of the camera).

Worth mentioning is the fact that emotions from facial expressions recognition is sometimes used in evaluation experiments in a non-automatic way. In their research on the role of humor in task-oriented dialogue, Morkes et al. [47] checked the participants' mirth reaction by counting how many times they smiled during the experiment. This method required only a video camera, as there was no emotion recognition system employed – however, it can also be a good complementary mean, especially in research related to humor.

7.4.3. *Multimodal detection*

As mentioned above, probably the best way to detect human emotions would be to create a complex system that works on multiple layers – i.e. textual, audio and visual. In recent few years some research projects tried to create multimodal emotion detectors – worth mentioning is the work by Fragopanagos and Taylor [48], who proposed neural network architecture able to handle the fusion of facial features, prosody and lexical content in speech. Also, other existing projects focus on joining two layers, mostly audio and visual.

Some of them base on artificial data (e.g. [49], [50]), while other prefer spontaneous affect displays (e.g. [51], [52]).

Needless to say, systems that would detect users' emotions in as complex and multilayer way as possible, would be of high importance in the field of HCI evaluation.

7.5. *Emotion Recognition Directly From the Brain*

All emotion recognition methods described in above sections have different drawbacks, but all share the same problem: in a way, they are indirect. In all of them, be it detection from textual, audio or visual layer, we can study only ways in which emotions are expressed, not emotions per se, as they "appear" only in our brains.

In nowadays science, thanks to the development of such fields as brain computer/machine interfaces (BCI/BMI), it has become possible to detect emotions directly from human brain. Apart from obviously related fields, such as psychology, new possibilities of neuroscience were especially welcomed in economy and marketing, as they can be easily used to evaluate products and check customer's reactions.

In this section we briefly investigate the possibilities of using certain methods of brain scanning in evaluations of chatterbots.

7.5.1. *Neuroeconomics and Neuromarketing*

Neuroeconomics can be seen as a convergence of psychology, economy and neuroscience. Its main area of interests is to analyze and understand economically relevant behaviour of potential customers [53]. In particular, current neuroeconomics research projects focus on such topics as customer's preferences [54], decision making [55], product choices [56], social interactions [57] or cooperative behaviour [58]. Most of these subfields are also of high importance in a field called "neuromarketing", which, comparing to neuroeconomics, is more closely related to applying neuroscientific methods to markets and marketing changes. Neuromarketing is often said to be focusing on finding so-called "buy button" in the brain, and, although quite simplified, this goal clearly shows the tendency in this field of science: we want to know why people like things and, subsequently, how to make things they like.

As science-fiction-like as it may sounds, this approach is not just a theory any more. In last few years, a number of neuromarketing agencies (BrightHouse in US, Neurosense and Neuroco in UK) have emerged, offering services in solving commercial problems. Most of these agencies focus on analyzing customer's reactions to TV commercials or billboards, which helps to find the most efficient way to advertise and sell a product. Neuromarketing scientists also focus on such topics as trust [59] or negotiations [60].

Therefore, evaluation methods proposed in this section can be seen as application of neuromarketing methods and approaches in the field of HCI. When we see chatterbots as products and users as customers, it becomes clear that the same (or similar) neuromarketing mechanisms must be working also in this case.

7.5.2. *Possibilities for HCI Evaluation*

Below we briefly summarize main methods of brain imaging. As the subject of our research is a chatterbot, we focus only on techniques that have been used or are possible to be used in the field of HCI or humor recognition. Needless to say, this section does not exhaust the subject and we are open for discussion on other possibilities of using neuroscience techniques in our research.

EEG

One of preferred method in this field is the use of an electroencephalograph (EEG) in order to detect brain responses to emotional stimuli [61].

The main basis of this approach relies on the knowledge that emotional states can be estimated basing on the asymmetry in the frontal brain activity in recorded EEG. Methods of analysing these asymmetries are summarized in the work of Coan and Allen [62]. Generally, most scientists today focus on the reduction in alpha band (8-13 Hz) activity, for it is experimentally proved that there is an inverse relationship between alpha activity and brain activation when experiencing emotions. In other words – when emotions are experienced, we can observe a reduction in alpha band activity, which can be measured using EEG [61].

Experiments with the use of EEG usually require participants to wear sets of electrodes (such as an EEG cap), connected to front head channels. The participants are then exposed to various emotion-inducing stimuli, such as videos or sounds inducing various emotional states [63]. Change of EEG power patterns are then analysed and compared for different emotional stimuli.

EEG caps and other devices used in brain computer/machine interaction are usually not very inconvenient and designed so that they do not disturb the experiment subjects.

Talking with a chatterbot in an EEG cap may sound ridiculous, but in fact it may be the best idea to detect users' emotive states, directly and noninvasively. Wearing the cap may be a little inconvenient, but it would allow us to save the users the trouble of evaluating the conversation by filling out a questionnaire. Also, as mentioned above, some emotions may only occur subconsciously, and not appear in the facial or textual layer. Therefore, this method seems relatively reliable and at least worth examining.

There are, however, some potential problems with the use of EEG in evaluation of our system. First, EEG signal records are often prone to contamination by noise due to the presence of devices in the surrounding that create electromagnetic interference. This problem could be solved by using Rutkowski's et al. empirical mode decomposition (EMD) – a new technique of decomposing EEG signals, proved successful in detecting emotive reactions to video stimuli [63].

The second problem is due to the fact that our systems are textual-input based. In experiments using EEG, the amount of movements from the participants has to be limited, since muscle work disturbs the readings. Therefore, typing utterances on a keyboard could seriously contaminate the records. One possibility to solve this problem would be

to use a method called "functional near-infrared spectroscopy" (fNIRS) – an emerging imaging technology measuring blood oxygenation levels in the brain [64].

fNIRS

In fNIRS deoxygenated and oxygenated haemoglobin are the main absorbers of near-infrared light in tissues, which makes them relevant markers for hemodynamic (dynamics of the blood flow) changes in the brain. In other words – this method allows us to check how much oxygen was used by the neurons, and thus – how much they were engaged in the process of conveying of information.

Therefore, as the method is not directly focused on power changes in the frontal lobe, it is robust against interferences such as those due to electromagnetic fields. It is also non-invasive and requires only wearing a hair band, which holds a fNIRS device placed on both sides of the forehead (see Figure 10).

These features make this method a good alternative for conventional EEG techniques, especially when computers or other electronic devices are involved. The idea of using fNIRS in HCI was proposed by Hirshfeld et al. [65]. In their experiments they used this method to classify different levels of mental workload during performance of a task with the computer. The average accuracy in the experiment was on the level of 83%, which is high enough to justify the use of fNIRS in HCI and makes it a promising technology to use also in our research.

Fig. 10. Experiment participant wearing fNIRS device [65]

Humor Detection – fMRI and fNIRS

While the section above describes the idea of emotive states recognition from the brain in general, in the case of our research, there is one more feature that is of high importance: humor. Detecting brain reactions to humorous stimuli is not a new idea – some research on this subject has been done by Goel and Dolan [66], Moran et al. [67] or Mobbs et al. [68]. In most projects, a method called "functional magnetic resonance imaging" (fMRI) was used. It measures the oxygen level dependent (BOLD) changes, related to neural activity in the human brain. This is quite similar to what fNIRS does. However, fMRI can

provide us with far more accurate and robust results, as it can also scan deep parts of brain.

Since the BOLD changes seem to be strictly related to appreciation of humor [67], this method could be used in our research on joking systems, in order to check how the system's humorous utterances were appreciated by humans.

Comparing to EEG, fMRI has high spatial resolution but relatively poor time resolution. EEG, however, directly measures electrical activity of the brain, while fMRI relies on the blood flow. Therefore, although technically challenging, simultaneous use of these two methods is possible [69] and should allow us to measure human's emotional states and his/her responses to humorous stimuli at the same time.

However, there is one literally big problem with fMRI - the size of the scanner. In case of EEG (cap) and fNIRS (device held by a hair band), there was not much inconvenience for the user and he/she could relatively freely perform other physical activities. This is impossible while being scanned in the fMRI device, which requires placing the subject in a strong magnetic field (see Figure 11).

This makes it impossible to perform a conversation with the chatterbot while being scanned by an fMRI device. Therefore, if we were to use this technology, we would have to limit the experiment to third person evaluators, to which chat logs of both non-humor- and humor-equipped systems would be read aloud when they are placed inside the device. This method, albeit not direct, would possibly allow us to check participants' response towards computer-generated humor stimuli, their appreciation and conveyed emotions at the same time.

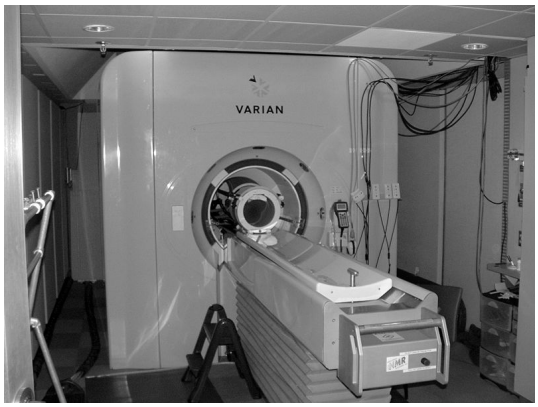

Fig. 11 Berkeley's 4T fMRI scanner (http://en.wikipedia.org/wiki/File:Varian4T.jpg)

8. Future Challenges

In this paper we proposed a methodology of chatterbot evaluation, focusing on the non-linguistic area. In the section **7** we described and discussed the evaluation methods, using the results of experiments from our previous research as examples. As the experiments have been actually conducted, it can be stated that the presented methods are at least applicable and can be used also in other research on non-task oriented dialogue systems.

Needless to say, the methods and methodology itself are not perfect and need to be improved. Some issues that should be addressed are described in below sections.

8.1.1. *Embodiment and Voice Recognition / Generation*

Since both systems evaluated in our experiments were text-based only, we did not include such issues as visual or audio quality of the systems' performance. However, many dialogue systems do have embodied versions and/or voice generation and recognition modules. In such cases, the quality of these aspects of interaction also has to be taken into consideration. In our methodology (see Figure 1) we mentioned these features in the linguistic features/technical quality section; however, the methods here need further discussion. Also, in the non-linguistic area of evaluation, such issues as embodiment should be taken into consideration, as they influence on user's impression about the system.

8.1.2. *Interlocutor's Identity*

One important question in both first and third person oriented evaluation methods is whether the users should know that their conversation partner is a computer system. In our experiments, the users did have that knowledge, contrary to the non-users evaluators. We believe that repeating the same experiments with opposite settings can give us an answer about the correlation between the knowledge of interlocutor identity and evaluation results. For example, when the users do know that it is a computer that talks to them and says jokes, this very fact can be appreciated and reflected in the evaluation. This may mean that it is more objective to hide the non-humanness of the partner before the evaluators – however, more experiments in this area are needed.

8.1.3. *Duration and Repetitiveness of Interaction*

In the experiments described above, interactions between the users and the systems were 10-turns long. This was partially caused by the fact that both our systems use the Internet to generate responses, and search engines, such as Goo or Google, could block our IP every time we do too many queries simultaneously. However, we think that it would be a good idea to conduct evaluation experiments also in the long run. The users should interact with the systems for a longer time, and the interaction should be repeated in order to check if the users' first impression was not just caused by the novelty of the system.

The questions: how long should one interaction be and how many times should it be repeated are open for further investigation.

8.1.4. *Individual Differences*

The fact that most of the methods described in this paper are not objective does not mean that they are useless. It is the client (user) who is going to buy and use our product, and thus his/her subjective impression is of highest importance. However, the subjectivity of evaluation can lead to a situation when, even if we construct a very sophisticated system, that will be evaluated highly by most of evaluators, we still cannot be sure that all users will like it, as there are individual differences that influence the assessment.

The best method to prevent such situations is to construct such a system that would adapt to the user's needs. In our research we focus on the role of humor in the conversation. Currently we are working on an emotive-analysis-based evolution of humor algorithm (its outline was presented in one of our earlier works [24]), which will allow the system to check user's reactions to particular jokes (using the ML-Ask system – see **6.3**) and on this basis build his/her sense of humor model. For example, if the user reacts with positive emotions to jokes concerning politics, the system can assume that this type of joke matches his/her sense of humor. In this manner, the longer the system would talk to the user, the more accurate "tags" of humor sense it can attach – and this, in effect, shall lead to more personalized, more individualized jokes that with a high probability would be appreciated by the user.

Needless to say, such individualization algorithm would also have to be evaluated, which would require proper questions to be included in the user-evaluation.

8.1.5. *Scales*

As discussed in **4**, in some of existing works (c.f. [25]) evaluation questionnaires do not include any scales and require the users to answer the questions freely. In our research though we decided to use the scales, as it is easier to interpret and compare numbers than written impressions. In both first and third person oriented experiments we used 5-point scales, with some descriptions added to each value (e.g. 1: No, I don't think so at all, 2: Not really, 3: Maybe a little, 4: Let's say I think so, 5: I think so). From the results we can state that for the needs of our research this was sufficient, but not necessarily the most adequate. We would like to see some research on measuring scales in non-linguistic areas of dialogue systems evaluation, so that the most appropriate set of scores could be chosen.

8.1.6. *User Simulation*

In some evaluation experiments of task-oriented dialogue systems user simulation techniques are used to evaluate the system's performance (e.g. [70], [71]). It may work well for linguistic areas of evaluation – however, when we are to explore the non-linguistic issues, such as the system's human-likeness or user's will to continue the dialogue, automatic simulation of user's behavior would be of little help in the first person oriented methods. However, it may work in the third person experiments, where the chat logs evaluated by non-user participants could be generated automatically, between the user simulating system and the system we want to evaluate. The main

condition here would be that the simulation should be good enough (close to human level), so that it would not disturb the evaluators, which could influence the results.

9. Conclusions

In above section we presented our evaluation methodology for the evaluation of non-linguistic (subjective) features of chatterbots. Most of the methods were tested in actual experiments, which means that they are at least applicable and should prove useful in the evaluation of HCI, in particular in its conversational layer. Also, the methods proposed as possible to apply seem useable and worth trying.

The methodology is obviously not perfect and still needs many improvements. One of the aims of this paper is to give a new impetus to the discussion on evaluation of chatterbots, especially in the non-linguistic area. It is high time that the gap in this field was finally bridged and a robust methodology was created, with universal methods that could be reused in the evaluation of HCI in general.

References

1. S. Gandhe and D. Traum, Creating Spoken Dialogue Characters from Corpora without Annotations, in *Proceedings of Interspeech-07* (Antwerp, Belgium, 2007)
2. J. Weizenbaum, ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1) (1966) pp. 36-45.
3. T. Winograd, *Understanding Natural Language* (New York Academic Press, 1972).
4. S. K. Feiner and K. R. McKeown, COMET: Generating coordinated multimedia explanations, in Proceedings of CHI'91, eds. S. P. Robertson, G. M. Olson and J. S. Olson (Reading, MA: Addison-Wesley 1991), pp. 449-450.
5. J. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. G. Martin, B. W. Miller, M. Poesio and D. R. Traum, The TRAINS project: A case study in building a conversational planning agent, *Journal of Experimental and Theoretical AI* 7 (1995) pp. 7-48. Also available as University of Rochester, Dept. of Computer Science TRAINS Technical Note 94-3.
6. R. W. Smith and S. A. Gordon, Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue, *Computational Linguistics*, *23*(1) (1997) pp. 141-168.
7. N. Reithinger, J. Alexandersson, T. Becker, A. Blocher, R. Engel, M. L¨ockelt, J. M¨uller, N. Pfleger, P. Poller, M. Streit and V. Tschernomas, Smartkom - adaptive and flexible multimodal access to multiple applications, in *Proceedings of ICMI 2003* (Vancouver, B.C., 2003).
8. Z. Callejas, R. López-Cózar, Implementing Modular Dialogue Systems: A Case Of Study, *ISCA Tutorial and Research Workshop on Applied Language Interaction in Distributed Environments* (2005).
9. A. Raux, B. Langner, D. Bohus, A. Black and M. Eskenazi, Let's Go Public! Taking a Spoken Dialog System to the Real World, in *Proceedings of Interspeech 2005 (Eurospeech)* (Lisbon, Portugal, 2005).
10. K. M. Kim, J. H. Hong and S. B. Cho, A semantic Bayesian network approach to retrieving information with intelligent conversational agents, *Information Processing & Management*,

Vol. 43,1 (2007), pp. 225-236.
11. O. Lemon, K. Georgila, J. Henderson and M. Stuttle, An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system, in *Proceedings of EACL (demo session) 2006* (Trento, Italy, 2006).
12. T. Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer (2001) http://mecab.sourceforge.net/
13. L. Dybkjær , N. O. Bernsen and W. Minker, Evaluation and usability of multimodal spoken language dialogue systems, *Speech Communication 43*(1–2) (2004), pp. 33–54.
14. A. Turing, Computing Machinery and Intelligence. *Mind 59*(236) (1950), pp. 433-460.
15. A. P. Saygin, I. Cicekli and V. Akman, Turing Test: 50 Years Later, *Minds and Machines, 10*(4) (2000), pp. 463–518.
16. S. Higuchi, R. Rzepka and K. Araki, 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)* (Honolulu, USA), pp. 382-390.
17. P. Dybala, M. Ptaszynski, S. Higuchi, R. Rzepka and K. Araki, Humor Prevails! - Implementing a Joke Generator into a Conversational System. In *Proceedings of the 21st Australasian Joint Conference on AI (AI-08)*, Wobcke, W. and Zhang, M. (eds), Auckland, New Zealand, 2008. Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI) Vol. 5360 (Springer Berlin & Heidelberg, 2008), pp. 214-225.
18. M. Ptaszynski, P. Dybala, S. Higuchi, R. Rzepka, and K. Araki, How to find love in the Internet? Applying Web mining to affect recognition from textual input. In *Proceedings of the 2008 Empirical Methods for Asian Languages Processing Workshop (EMALP'08) at The Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI'08)* (Hanoi, Vietnam, 2008), pp. 67-79.
19. M. 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, Poland, 2006).
20. M. Ptaszynski, P. Dybala, S. Higuchi, R. Rzepka and K. Araki, Affect as Information about Users' Attitudes to Conversational Agents, In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'08), Second International Workshop on Human Aspects in Ambient Intelligence (HAI'08)* (Sydney, Australia, 2008), pp. 459-500.
21. J.A. Russell, A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6) (1980), pp. 1161-1178.
22. A. Nakamura, *Kanjo hyogen jiten (Dictionary of Emotive Expressions)* (in Japanese), (Tokyodo Publishing, Tokyo, Japan, 1993).
23. 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.
24. P. Dybala, M. Ptaszynski, R. Rzepka and K. Araki, Humorized Computational Intelligence - towards User-Adapted Systems with a Sense of Humor, In *Proceedings of the EvoStar 2009 Conference, EvoWorkshops*, M. Giacobini et al. (Eds.). Springer-Verlag Lecture Notes in Computer Science (LNCS), Vol. 5484 (Springer Berlin & Heidelberg, 2009), pp. 452–461.
25. N.O. Bernsen and L. Dybkjær, Evaluation of spoken multimodal conversation. In *Proceedings of the 6th International Conference on Multimodal Interaction (ICMI'04)*, (New York: ACM Press, 2004), pp. 38-45.
26. P. Dybala, M. Ptaszynski, R. Rzepka and K. Araki, Activating Humans with Humor - A Dialogue System that Users Want to Interact With, *IEICE Transactions on Information and Systems Journal, Special Issue on Natural Language Processing and its Applications, Vol.E92-D, No.12*, (December 2009), pp. 2394-2401.

27. M. Selting, Emphatic speech style - with special focus on the prosodic signalling of heightened emotive involvement in conversation, *J. Pragmatics*, *22* (3–4) (1994), pp. 375–408.
28. M.H. Goodwin and C. Goodwin, Emotion within situated activity, *Communication: An Arena of Development*, eds. N. Budwig, I. C. Uzgiris, J. V. Wertsch, (Ablex, 2000), pp. 33-54.
29. C. Yu, P. M. Aoki and A. Woodruff, Detecting user engagement in everyday conversations. In *Proceedings of 8th International Conference on Spoken Language (ICSLP 2004)*, (2004, Jeju Island, Korea), pp. 1329-1332.
30. O. W. Kwon, K. Chan, J. Hao and T.W. Lee, Emotion Recognition by Speech Signals, In *Proceedings of EUROSPEECH* (2003, Geneva, Switzerland), pp. 125-128.
31. A. Austermann, N. Esau, L. Kleinjohann and B. Kleinjohann, Prosody based emotion recognition for MEXI, in *Proceedings of Int. Conf. Intelligent Ro‐bots and Systems (IROS 2005)* (2005, Edmonton, Alberta, Canada), pp. 1138‐1144.
32. L. Devillers and I. Vasilescu, Real‐Life Emotions Detection with Lexical and Paralinguistic Cues on Human‐Human Call Center Dialogs, in *Proceedings of Int. Conf. on Spoken Language Processing* (Pittsburgh, USA, 2006).
33. C. M. Lee and S. S. Narayanan, Toward detecting emotions in spoken dialogs. *IEEE Tran. Speech and Audio Processing*, *Vol. 13*(2) (2005), pp. 293‐303.
34. D. Neiberg, K. Elenius and K. Laskowski, Emotion Recognition in Spontaneous Speech Using GMM, in *Proceedings of Int. Conf. on Spoken Language Processing* (Pittsburgh, USA, 2006), pp. 809‐812.
35. C. Blouin and V. Maffiolo, A study on the automatic detection and characterization of emotion in a voice service context, *in Proceedings of Interspeech* (2005, Lisbon), pp. 469-472.
36. D. J. Litman and K. Forbes‐Riley, Predicting Student Emotions in Computer‐Human Tutoring Dialogues, in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)* (2004, East Stroudsburg, USA), pp. 352-359.
37. P. Ekman, An Argument for Basic Emotions, *Cognition and Emotion*, 6 (1992), pp. 169-200.
38. Y. L. Tian, T. Kanade and J. F. Cohn, Facial expression analysis, In: *Handbook of Face Recognition*, Li S Z and Jain A K (Eds.) (Springer, New York, USA, 2005), pp. 247-276.
39. M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. Fasel and J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in *Proceedings of Int. Conf. on Automatic Face and Gesture Recognition* (2006, Southampton, England), pp. 223-230.
40. J. F. Cohn and K. L. Schmidt, The timing of Facial Motion in Posed and Spontaneous Smiles, *International Journal of Wavelets, Multiresolution and Information Processing*, *2* (2004), pp. 1-12.
41. A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon and B. J. Theobald, The painful face: pain expression recognition using active appearance models, in *Proceedings of Int'l Conf. Multimodal Interfaces* (2007), pp. 9-14
42. M. Yeasin, B. Bullot and R. Sharma, Recognition of facial expressions and measurement of levels of interest from video, *IEEE Trans. On Multimedia*, *Vol.8,* No. 3 (June 2006), pp. 500-507.
43. M. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions, in *Proceedings of Int. Conf. on Multimedia Interfaces* (2006), pp. 162-170.
44. Y. Chang, C. Hu, R. Feris and M. Turk, Manifold based analysis of facial expression, *J. Image and Vision Computing*, *Vol. 24,* No.6 (2006), pp. 605-614.
45. G. C. Littlewort, M. S. Bartlett and K. Lee, Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain, in *Proceedings of Int'l Conf. Multimodal Interfaces* (2007), pp. 15-21

46. J. Xiao, T. Moriyama, T. Kanade and J. F. Cohn, Robust full-motion recovery of head by dynamic templates and re-registration techniques, *Int. J. Imaging Systems and Technology*, *Vol. 13*, No.1 (2003), pp. 85-94.
47. J. Morkes, H. K. Kernal and C. Nass, Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of srct theory, *Human-Computer Interaction*, *14*(4) (1999), pp. 395-435.
48. F. Fragopanagos and J. G. Taylor, Emotion recognition in human–computer interaction, *Neural Networks*, *Vol. 18* (2005), pp.389-405.
49. C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann and S. Narayanan, Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, in *Proceedings of Int. Conf. Multimodal Interfaces*, (2004, Pennsylvania, USA) pp. 205-211.
50. Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth and S. Levinson, Audio-visual Affect Recognition, *IEEE Transactions on Multimedia*, *Vol. 9*, No. 2 (February 2007), pp. 424-428.
51. G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A, Paouzaiou and K. Karpouzis, Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition, in *Proceedings of Int. Conf. on Multimodal Interfaces* (2006), pp. 146-154.
52. S. Petridis and M. Pantic, Audiovisual discrimination between laughter and speech, in *Proceedings of Int'l Conf. Acoustics, Speech, and Signal Processing* (2008, Las Vegas, USA), pp. 5117-5120.
53. C. F. Camerer, G. Loewenstein and D. Prelec, Neuroeconomics: How Neuroscience can Inform Economics, Working Paper, UCLA Department of Economics, Levine's Bibliography, (2003).
54. S. M. McClure, J. Li, D. Tomlin, K. S. Cypert, L. Montague and P. R. Montague, Neural correlates of behavioral preference for culturally familiar drinks, *Neuron* 44 (2004), pp. 379-387.
55. M. Deppe, W. Schwindt, H. Kugel, H. Plassmann and P. Kenning, Nonlinear responses within the medial prefrontal cortex reveal when specific implicit information influences economic decision making, *J. Neuroimaging 15* (2005), pp. 171-182.
56. T. Ambler, S. Braeutigam, J. Stins, S. P. Rose and S. J. Swithenby, Salience and choice: neural correlates of shopping decisions, *Psychol. Market.* 21 (2004), pp. 247-266.
57. S. Erk, M. Spitzer, A. P. Wunderlich, L. Galley and H. Walter, Cultural objects modulate reward circuitry, *Neuroreport 13* (2002), pp. 2499-2503.
58. K. McCabe, D. Houser, L. Ryan, V. Smith and T. Trouard, A functional imaging study of cooperation in two-person reciprocal exchange, in *Proceedings of Natl. Acad. Sci. U.S.A.* 98 (2001), pp. 11832-11835.
59. B. King-Casas, D. Tomlin, C, Anen, C. F. Cemerer, S. R. Quartz and P. R. Montague, Getting to know you: reputation and trust in a two-person economic exchange, *Science 308* (2005), pp. 78-83.
60. A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom and J. D. Cohen, The neural basis of economic decision-making in the ultimatum game, *Science 300* (2003), pp. 1755-1758.
61. C. P. Niemic, Studies of Emotion: A Theoretical and Empirical Review of Psychophysiological Studies of Emotion, *Journal of Undergraduate Research, v. 1, no. 1* (2002), pp. 15-18.
62. J. A. Coan and J. J. B. Allen, Frontal EEG asymmetry as a moderator and mediator of emotion, *Biological Psychology*, *67* (2004), pp. 7-49.
63. T. M. Rutkowski, A. Cichocki, A. L. Ralescu and D. P. Mandic, Emotional states estimation from multichannel EEG maps, In *Advances in Cognitive Neurodynamics ICCN 2007*

Proceedings of the International Conference on Cognitive Neurodynamics, eds. R. Wang, F. Gu, and E. Shen, *Neuroscience* (Springer Berlin & Heidelberg, 2009), pp. 695-698.
64. B. Chance, E. Anday, S. Nioka, S. Zhou, L. Hong, K. Worden, C. Li, T. Murray, Y. Ovetsky and R. Thomas, A novel method for fast imaging of brain function, non-invasively, with light, *Optics Express*, *10* (2) (1988), pp. 411-423.
65. L. M. Hirshfield, A. Girouard, E. T. Solovey, R. J. K. Jacob, A. Sassaroli, Y. Tong and S. Fantini, Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy, Presented on ACM UIST 2007 Symposium on User Interface Software and Technology, Poster Paper.
66. V. Goel and R. J. Dolan, The functional anatomy of humor: segregating cognitive and affective components, *Nat. Neurosci.* 4 (2001), pp. 237-38.
67. J. M. Moran, G. S. Wig, R. B. Adams, P. Janata and W. M. Kelley, Neural correlates of humor detection and appreciation, *Neuroimage* 21 (2004), pp. 1055-1060.
68. D. Mobbs, M. D. Greicius, E. Abdel-Azim, V. Menon and A. L. Reiss, Humor modulates the mesolimbic reward centers, *Neuron 40* (5) (2003), pp. 1041-1048.
69. R. I. Goldman, J. M. Stern, J. Engel and M. S. Cohen, Simultaneous EEG and fMRI of the alpha rhythm, *Neuroreport 13* (2002), pp. 2487-2492.
70. W. Eckert, E. Levin and R. Pieraccini, User modeling for spoken dialogue system evaluation, in *Proceedings of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU '97) (1997), pp. 80-87.
71. J. Schatzmann, K. Georgila and S. Young, Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems, in *Proceedings of 6th SIGdial Workshop on Discourse and Dialogue* (2005, Lisbon, Portugal), pp. 45-54.