Humor intelligence for virtual agents

Andreea I. Niculescu and Rafael E. Banchs[1]

Abstract Humor is pervasive in human social relationships and one of the most common ways to induce positive affect in others. Research studies have shown that innocent humor increases likeability, boosts trust, reduces tension, encourages creativity and improves teamwork. In this paper, we present a study focusing on deploying humor in interaction with a virtual agent. 25 participants evaluated the logs of conversations exchanged between a human user and two virtual agents acting as tour guides. Even though answers were equal in terms of content delivered, one agent used humorous statements to respond the queries while the other agent presented the content in a neutral way. To create answers with a humorous effect we combined information extracted from various websites focusing on tourist fun facts, pun and jokes collections. Results showed that our manipulation was successful, i.e. the humorous agent was indeed perceived as being significantly funnier. Additionally, the agent was perceived as delivering more interesting answers as compared with its counterpart. Further, participants showed statistically significant preferences towards the humorous agent when asked to choose between the agents. As such, we believe that using humor in interaction with virtual agents increases the agent likeability and possibly contributes towards a better user experience.

1 Introduction

Humor is a pervasive phenomenon in human society being described as the natural tendency to create laughter and good mood [16]. Research has shown that humor has biological roots and plays a beneficial role in a variety of social functions, such as reducing stress, aiding in education or defining gender identity [5].

Over the time, humor has been intensively investigated in many areas, such as psychology, sociology, biology, linguistics or computer science (CS). In the natural language processing (NLP) and artificial intelligence (AI) communities for example, humor is an established field with own research agenda. In contrast, in the human computer interaction (HCI) humor appears to be a rather neglected research topic, despite that -as a CS discipline focusing on interaction – it could greatly benefit from its deployment.

[1] Andreea I. Niculescu and Rafael E. Banchs
Institute for Institute for Infocomm Research, A*STAR
1 Fusionopolis Way #21-01 Connexis South Tower - Singapore 138632
e-mail: {andreea-n, rembachs}@i2r.a-star.edu.sg

One reason is that HCI traditionally focus on interfaces meant to increase task performance on one side, and minimize task duration, learning time and error rate, on the other side. Since the use of humour would distract users from their tasks increasing the total completion time, it would contradict HCI policies of maximizing efficiency in interaction [Niculescu et al., 2013].

However, in the future we can expect HCI to become less goal directed not only in entrainment computing, but also in our ordinary daily life. Technology moves slowly from working environment to our living room where computers and artificial entities are becoming social actors [12]. Therefore, it is important for interaction designers working in applied AI to take into account the social aspect of interaction between humans and computers where humor could become an influential ingredient towards technology acceptance and overall user satisfaction.

As such, in this paper, we are exploring the use of humor in interaction with a virtual tour guide. Currently, our tour guide combines task-oriented dialogues with Chabot functionalities: it offers information about Singapore, as well as informal chatting.

To explore the impact of humor in interaction with the agents, we handcrafted conversation scripts with humorous answers and asked test participants to evaluate them against scripts with neutral answers. Since the answers were manually generated by our team, there was no external validation on whether they were indeed perceived as humorous.

Thus, the goal of our experiment was twofold: 1) first, the study is aiming to validate the handcrafted humor data set; 2) second, the study tries to verify our hypothesis concerning the participants' preference towards the humorous agent; according to this hypothesis, we expect participants to perceive the information delivered by the humorous agent as more useful and the interaction more interesting.

2 Related work

A handful of researchers explored the use of humor in interaction with virtual entities. The study by Morke et al. [8] found that participants who received humorous comments from a computer rated the system significantly more cooperative, more likable and more competent. During the experiment those who received the humorous comments also smiled and laughed more displaying a more social behaviour as compared to those who didn't receive such comments.

Another study by Huan and Szafir [5] investigated humor in computer mediated learning. Their results showed that using humor significantly increased the instructor likeness, regardless if this was a human or robot.

The study by Niculescu et al. [9] showed that humour in interactive task with a social robot receptionist increase the users' likeliness towards the robot's speaking style and personality, as well as towards the overall task enjoyment.

Also, the study performed by Dybala et al. [4] proved that users interacting with conversational agents rated the humorous agent as being more funny and likeable

and more human-like. Comparing the two agents - humorous vs. non-humorous - the authors found out that better ratings were given to the humorous agent.

Humor can be also used to help a Chabot system recovering from errors: a study by Niculescu and Banchs [11] showed that using humor in situations when the system is unable to retrieve the correct answer may prompt the user to reformulate the query, thus helping the system to recover from errors.

Last but not least, the study by Barbu et al. [4] found that a humorous virtual receptionist may foster user participation in social conversations by using jokes.

All these experiments demonstrate that humor can have positive effects in HCI enhancing the overall user experience and keeping the users engaged in the interaction with the system.

3 The virtual tour guide

Our experiment is based on virtual tour guide that offers touristic information about Singapore. At the core, the tour guide is a multimodal dialogue system based on a client-server architecture. The system is composed of two modules operating in cascade: a rule-based module accessing information from two databases - a handcrafted database and a database composed of resources automatically extracted from web directories – and a data driven module accessing information from an index based on Wikipedia articles. A learning module complements the architecture enabling the system to learn from users' answers[2]. The overall system architecture is presented in the figure 1.

3.1 Databases

The first data base – the handcrafted database - is created by adding manually responses about relevant tourist spots in Singapore. It includes a total of 75 locations of interests concerning museums, theatres, temples, historical buildings, parks and heritage sites. The automatically collected data resources – the second database – come from crawling web directories about Singapore.

The database contains about 8000 entries concerning shopping malls, eateries, hotels, restaurant recommendations, transportation, stores etc. The system would access this database only if it fails to find an answer in the first database.

If the system fails to find an answer in the second database as well, the user input is passed to the data driven module. This module is implemented by an example-based question answering system. While the databases deal with queries that are specific to touristic attractions and other venue locations, the index contains a collection of question-answer pairs on more general information about Singapore.

The index data collection was automatically populated by crawling and processing Wikipedia pages related to Singapore [7]. Each index pair is matched with

[2] At the moment, this module is disabled and under development.

a given input query to compute the contextual similarity between them [12]. If an index pair has a higher similarity score compared to other examples, its answer is considered as first candidate for the system response.

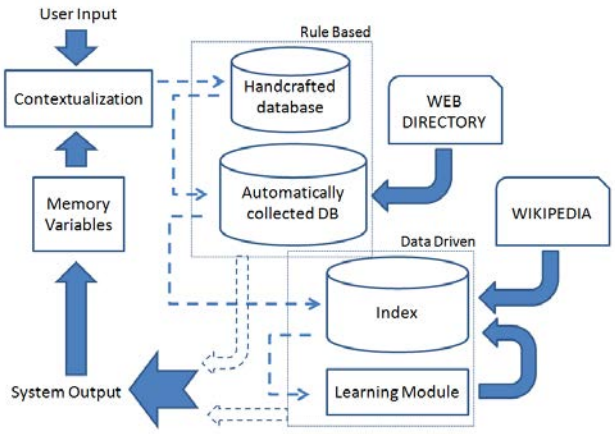

Figure 1 Overview architecture SARA

3.2 Multimodal dialogue system

Server side

As suggested in the previous paragraph, for the dialogue manager (DM) implemented in our system we chose two different strategies: a rule-based approach and an example based approach. The rule base approach uses a set of manually defined heuristics to determine an appropriate answer on the system site. For the example-based approach, we use Lucene[3], a special library meant for developing search engines.

We use the rule-based approach for handling task-oriented queries while the example-based approach focuses on more general questions.

Concerning the natural language understanding (NLU) module and dialogue topic tracking, the system uses a hybrid approach. The user input is transformed into a semantic representation using rules and statistical models. These are built based on data collected for the Singaporean touristic domain.

We used around 40 hours of human-human dialogue to train the system. The data was collected in both English and Chinese[3][4] as dialogues exchanges between visitors and tour guides. Dialogues were further manually annotated on several sematic levels, such as word, utterance, and dialogue segment.

---

[3] http://lucene.apache.org/

[4] For our current experiment, only the English version was used

To deliver the answer to the user, a natural language generation (NLG) module is being used. The NLG module of the rule-based dialogue manager uses a template-based approach.

For the data driven approach, no NLG component is used: the answers are provided exactly as extracted from Wikipedia pages. Once the answer is generated, it is further sent to the web client for image/map/web display and text-to-speech (TTS) generation.

The system main components are linked together using Apollo, a pluggable dialogue platform allowing the interconnection and control of the different independent components, as shown in figure 3 [6].

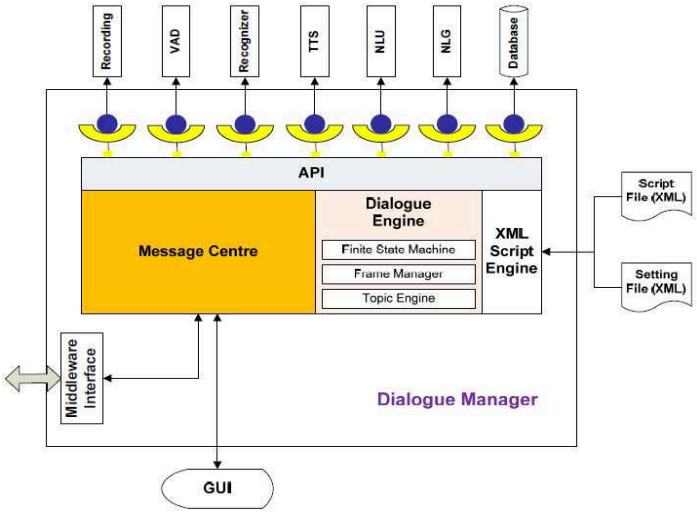

Figure 2 Apollo architecture

Client side

On the client side, the user interacts with a web browser. The web-based client interface shown in figure 3 has several components. From top to bottom, these components are: the avatar, the text input field and the response.

The avatar, as seen in the top left of the picture, is used to provide a spoken response to the user by reading out the text returned by Apollo. The user can input his query in the upper field and receives the answer in the form of a spoken answer (by the avatar), as well as written as text – in figure 3, both question and the answer prompted by the system can be visualized in the black response area. The user can use typing or speech to input the query.

To input a query, the user needs to press the grey button located on the right-side of the input field. The web interface contains at the bottom additional information related to the query, such as websites related to the query, maps, traveling direction

etc. In figure 3, the user is presented with the website of the Singapore Flyer which is referenced in the question. This webpage is retrieved from the database and can be configured in the server-side scripts.

In terms of content, the system provides answers about local sightseeing, restaurant recommendations, travel advice based on the current user location, i.e. using Google location services, can make online reservation and send SMS for confirmation.

Additionally, the system understands questions in context, i.e. the system is able to related the question to the previous one shown in figure 3 – "there" stands for "Singapore Flyer".

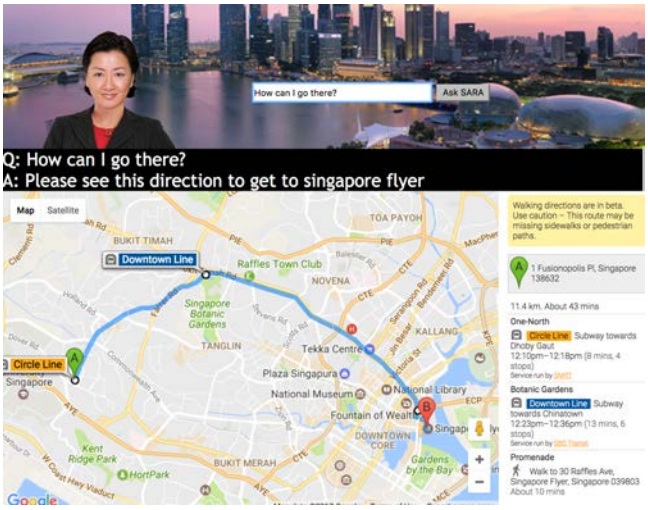

Figure 3 – System's web client

4 Methods

In order to carry out our experiment, we prepared 2 sets of conversation logs (set A and set B). Each set contained 3 different dialogues scripts (1, 2, 3). The participants were first introduced to the system and had the opportunity to play around with it, i.e. to ask their own questions and test its functionalities. Further, they were asked to read the scripts and to fill a questionnaire form after each log they read. The exposure to the log conversations was randomized between the agents to avoid biases.

4.1 Conversation logs

Each dialogue script was composed by 12 questions & answers (Q&A) pairs. The pairs covered greeting exchanges, chitchat *(What is your name? Do you*

like Chinese food?), personal questions *(Do you love me? Would you marry me?),* focused questions about locations of interest, stores, restaurant recommendations, direction advise and general purpose questions about Singapore (*How is the weather in Singapore? What is the exchange rate? Etc*.) Additionally, the logs contained questions asked in context, i.e. questions requesting anaphora resolution, as in the example shown in figure 3. We include in the scripts also questions that were unknown to the system, i.e. questions for those typically no answers could be found in the database.

Type: Question	**Set A** **Tyra – answer**	**Set B** **Sara answer**
Chat: "What is your job?"	"I am a tour guide!"	"I am a CIA agent disguised as tour guide! "
Chat: "What do you do?**"**	"I am a tour guide"	"I am painting my nails while waiting for people to ask me question about Singapore!"
Personal: "Do you love me ?"	"Sorry, I am not sure about that!" (*agent has no understanding of the question*)	"Ah, you are looking for love in all the wrong places!" (*agent has no understanding; use word association to create an answer)*
General: "How is the weather in Singapore?"	"Singapore has uniform temperature and pressure m high humidity, and abundant rainfall" *[Link on weather report of the day]*	"Usually hot, hot, hot like chilli peppers!" *[Link on weather report of the day]*
General: "What do you know about Singapore?"	"Singapore, officially the Republic of Singapore, is a sovereign city-state and island country in South East Asia"	"I know the most important things: don't walk around naked, don't spit, don't litter, don't carry around durians and always remember to flush the toilet"
General (more focused): "Are there any nice beaches?"	"In Sentosa, there are many nice beaches, such as Siloso or Palawan beach"	"Oh, well I never go to the beach because of skin cancer and all, but in Sentosa, there are many nice beaches such as Siloso or Palawan beach "
Focused (question in context): "What are the opening hours [of Underwater World]?"	"Unfortunately, the aquarium has closed down. "	"Unfortunately, the aquarium has closed down. I heard, after closing down they invited everyone to a sea BBQ party."

Table 1 Q& A examples from the logs used during the experiment

Participants were told the first set (A) came from a virtual agent called Tyra while the second set (B) was stemming from an agent named Sara. The difference between the Tyra log set was made at the answer formulation level: while the user questions for both Sara and Tyra were identical, the answer Tyra gave were formulated in a neutral form; in contrast, Sara's answers contained jokes and were usually formulated in a humorous way – see table 1 for a few examples. Both Sara and Tyra delivered the same amount of information in terms of content.

Tyra's logs were extracted from real conversations between users and system exchanged during past testing sessions. Sara's logs were adapted to follow a humorous paradigm: her statements were chosen from websites containing fun facts about Singapore [1] [14], funny puns [13] and humorous answer collections [3]. Each Sara script log included between 6 and 9 humorous statements.

4.1 Questionnaire

The questionnaire form comprised 3 sections, one for each log script. Each section was divided in 2 parts, one for each virtual agent. Participants filled in their response concerning the answer usefulness, whether the agent was funny, the conversation interesting and how much they liked the agent in that particular script. The questions were formulated as statements, i.e. "I liked the agent in this dialogue conversation" on a 5 point Likert scale from "*strongly disagree*" to "*strongly agree*". At the end of each section and after reading the scripts of both agents, were asked to choose which agent would they prefer to talk to. Additionally, they could leave comments, if any.

5 Results

The questionnaire allowed us to gather both another statistical and qualitative data in the form of comments.

5.1 Demographics

A total of 25 persons performed the test, 13 women and 12 men with ages between 19-51. They were chosen from staff and students working in our department. Good English knowledge was compulsory for participation.

The majority (19 - 76%) was composed by: local Singaporeans & long term residents (13) and Asians (6) from China, Malaysia, Myanmar and Vietnam. However, we also had participants from Colombia (2), USA (1), Romania (1) and Russia (2).

More than half of the participants (13 – 52%) were below 30 years old; 6 were between 30-40, 5 were between 40-50 and 1was between 50-60.

5.2 Quantitative data

As a Kolmogorov–Smirnov test for normality run on our data proved to be negative, we used a Wilcoxon signed-rank test. The analysis showed that SARA was significantly perceived as being funnier and more interesting to interact with as compared to Tyra across. This finding is valid across all three scripts – see table 2 for calculated Z asymptotic, median & median rank and p-values.

Script 1 N=25	**Answer is useful**	**Agent is funny**	**Conversation is interesting**	**I like the agent in this conversation**
Z	-1.748[b]	-4.086[c]	-2.942[c]	-1.882[c]
Asymp. Sig (2-tailed)	.080	**.000**	**.003**	.060
Sara: Mean rank+/Median	8.50 / 4	**12.30 / 4**	**8.46 / 4**	9.42 / 4
Tyra: Mean rank+/Median	4.56 / 4	5.50 / 2	5.00 / 3	8.00 / 4
Script 2 N=25	**Answer is useful**	**Agent is funny**	**Conversation is interesting**	**I like the agent in this conversation**
Z	-1.508[b]	-4.061[c]	-3.175[c]	-1.384[c]
Asymp. Sig (2-tailed)	.132	**.000**	**.001**	.166
Sara: Mean rank+/Median	4.00 / 4	**11.83 / 4**	**9.12 / 4**	9.63 /4
Tyra: Mean rank+/Median	4.67 /5	4.50 / 2	16.00 /3	9.25 /3
Script 3 N=25	**Answer is useful**	**Agent is funny**	**Conversation is interesting**	**I like the agent in this conversation**
Z	-1.748[b]	-4.086[c]	-2.942[c]	-1882[c]
Asymp. Sig (2-tailed)	.577	**.000**	**.005**	**.013**
Sara: Mean rank+/Median	6.40 /4	**11.33 /4**	**8.00 /4**	**9.33 /4**
Tyra: Mean rank+/Median	6.57 /4	4.50 /2	4.50 / 3	10.33 /3

Table 2 Results of the Wilcoxon Signed ranked test for the 3 scripts - b. based on positive ranks /c. based on negative ranks

In terms of content usefulness, there were no significant differences between the agents - thus infirming our initial hypothesis. The explanation is that, one side the answers were formulated equal in terms of content –as such, the humor did not appear to contribute towards answer's usefulness.

On the other side, the humor effect might not have had the strong effect that we hoped for. Also, our number of participants was relatively low to enable strong and significant statistical differences.

Also, there was no statistically significant difference between Sara and Tyra concerning the agent likeliness except for script 3 – where Sara was significantly more liked by participants as compared to Tyra. The reason for this preference might have been related to the agent's humorous answers: we found significant correlations for script 3 between liking Sara and the agent capability of being funny ($r=.81$, $p<.001$) and offering an interesting conversation ($r=0.78$, $p<.001$).

Concerning their preference for Sara or Tyra, from a total of 75 responses along the three different tests (25x3) 51 participants chose the agent Sara while 24 decided for Tyra – see figure 4. A Binomial test found the proportion of participants choosing Sara to be significantly higher (p=.002) as compared with the proportion of voters for Tyra – the proportion was measured at an expected .50 level.

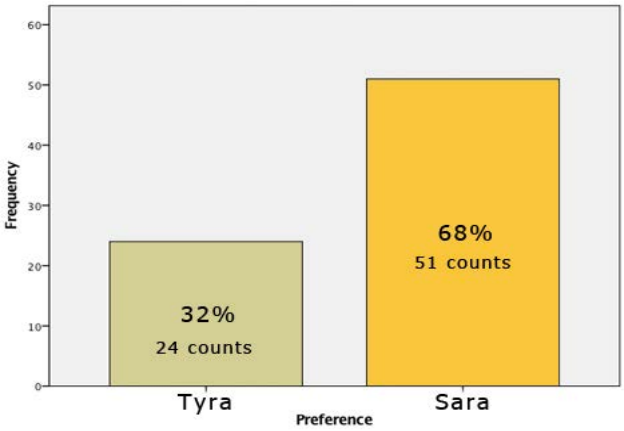

Figure 4. Total percentage difference between agents

5.3 Qualitative data

The questionnaire allowed us to collect a total of 31 comments from our participants. Since comments were optional not all participants choose to do so.

A common trend among the comments was that Sara was indeed perceived as funnier and more entertaining as comparing to Tyra. Tyra was often considered "*bland*", "*boring*" and "*unnatural*". Sara however, was perceived not equally funny in all scripts, e.g. some participants commented for example that script 3 was more entertaining as compared with script 2 – a fact that was also confirmed by our statistical analysis.

A foreign participant suggested the use of "*lah*", as "*it might sound funny*". "*Lah*" is a local spoken feature commonly used at the end of a sentence to emphasize content.

Another two participants criticized Sara for being sometimes too "*negative*" and "*blunt*" in her statements while three other participants seemed to have difficulties understanding some of the agent's jokes and thought the agent gave them a wrong answer.

Some local participants indicated that Sara's statements could be sometimes perceived as "*offensive*", "*negative*", "*sarcastic*" or even "*rude*": "*Sara is funny/sarcastic, but I am not sure if everyone find her like that. There may be some cultural/age determined sensitivity*" wrote a local participant.

In spite of that, Sara appeared to have more "*human touch*" as compared to the "*mechanical*" Tyra: "*You feel like you are speaking to a real person …*" wrote on of the participants.

Participants also expressed the concerns that humor despite being a nice way to complement the information might distract the user's attention from their search. Also, some participants commented that some of Sara's answers appeared to be too long due to the additional jokes, thus increasing the total interaction time.

6 Conclusions

The results of our experiment demonstrated that our humor manipulation was successful: the agent Sara was perceived significantly funnier as Tyra.

Further, our experiment showed that participants perceived the humorous agent engaging users in more interesting conversations. Also, participants seemed to prefer the humorous Sara over the more neutral Tyra. All these results suggest that humor might have beneficial effects when designed to support the interaction with virtual agents.

However, designing humorous responses is not an easy task: as pointed out by our participants' comments, humor effects are highly context dependent and subjected to cultural constrains. More traditional cultures, such as the Asian cultures for example, might be less open to use humor in formal situations or in situation of status inequality between individuals.

Personal preference and timing play also an important role on whether humor has the intended effect: no everyone has the same taste for jokes or is - at a given moment - in the right mood for jokes; also, some people might be in a hurry looking for a short, straight, informational answers, thus listing to jokes might be counterproductive for their goal.

Additionally, having a good command of the language spoken/written in interaction with the agent helps understanding the humorous intention.

Since Chabot systems – once available online - can be virtually accessed by anyone, having all these differences accounted for within a single system is a huge challenge that future research needs to address.

In the future, we plan to continue our work on implementing humor in interaction with our agent Sara by automatizing the responses created from a handcrafted database of humorous answers and jokes. The choice of answer will be based on the validated preferences from large test users pool. Further, we plan testing our humorous agent in direct interaction with users to detect best algorithmic strategies to complement the interaction with humor in the appropriate context.

Acknowledgments

We would like to thank to our 25 test participants who spent time and effort helping us with this experiment.

References

1. Avakian, T. Travel tips: 16 odd things that are illegal in Singapore. Online available: http://www.stuff.co.nz/travel/destinations/asia/70915066. Retrieved on 22.04.2018
2. Babu, S., Schmugge, S., Barnes, T., Hodges, L. F. What would you like to talk about? An evaluation of social conversations with a virtual receptionist. In Proceeding of Intelligent Virtual Agents, pages 169–180. Springer, LNCS Series (2006)
3. Carfi, J., Cliff C. Brilliant Answer for Everyday Questions, Carle & Carfi Publishing, Los Angeles CA (2012)
4. Dybala, P., Ptaszynski, M., Rzepka, R., Araki, K. Humoroids: Conversational Agents that Induce Positive Emotions with Humour. In Proceedings of 8th Int. Conf. on Autonomous Agents and Multi agent Systems (AAMAS), 1171–1172, (2009)
5. Huang, C.M., Szafir, D. No joke: Examining the Use of Humor in Computer mediated learning. Unpublished material, (2001)
6. Jiang, D. R., Tan, Y.K., Kumar Limbu, D., Li H. Component pluggable dialogue framework and its application to social robots, in Proc. Int'l Workshop on Spoken Language Dialog Systems (2012)
7. Kim, S., Banchs, R. E., Li, H. Wikipedia-based Kernels for Dialogue Topic Tracking. In Proceedings of ICASSP, (2014)
8. Morkes, J. Kernal, H.K., Nass, C. Effects of humor in task-oriented human-computer interaction and computer-mediated communication: a direct test of SRCT theory. Human-Computer Interaction, 14(4):395–435, (1999)
9. Niculescu, A.I., van Dijk, B., Nijholt, A., Li. H., See, S.L. Making social robots more attractive the effects of voice pitch, humor and empathy. International journal of social robotics, 5 (2): 171-191, (2013)
10. Niculescu, A.I., Yeo, K.H., D'Haro, L. F., Kim, S., Jiang, R., Banchs, R. E. Design and Evaluation of a Conversational Agent for the Touristic Domain. In Proceedings of Asia Pacific Signal and Information Processing Association (APSIPA), (2014)
11. Niculescu, A.I., Banchs, R.E. Strategies to cope with errors in human machine spoken interactions: using chatbots as back-off mechanism for task-oriented dialogues. In Proceedings of ERRARE, Errors by Humans and Machine in multimedia, multimodal and multilingual data processing, (2015)
12. Nijholt, A., Stock, O., Dix, A., Morkes, J.: Humour modeling in the interface. In: CHI'03 Extended Abstracts on Human Factors in Computing Systems. pp. 1050-1051. ACM (2003)
13. Salton, G., Wong, A., Yang, C. S. A vector space model for automatic indexing. Commun. ACM **18** (11), 613-620, (1975)
14. Pun of the day. Online available: http://www.punoftheday.com/cgi-bin/disppuns.pl?ord=F 2016. Retrieved on 22.04.2018
15. The Fact File: Interesting Facts about Singapore. Online available: http://thefactfile.org/singapore-facts/. Retrieved on 22.04.2018
16. Wikipedia Humor. Online available: https://en.wikipedia.org/wiki/Humour. Retrieved on19.02.2017.