2017 26th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
Lisbon, Portugal, Aug 28 - Sept 1, 2017.

Adapting a Robot's Linguistic Style Based on Socially-Aware Reinforcement Learning

Hannes Ritschel, Tobias Baur and Elisabeth André

Abstract— **When looking at Socially Interactive Robots, adaptation to the user's preferences plays an important role in today's Human-Robot Interaction to keep interaction interesting and engaging over a long period of time. Findings indicate an increase in user engagement for robots with adaptive behavior and personality, but also that it depends on the task context whether a similar or opposing robot personality is preferred. We present an approach based on Reinforcement Learning, which gets its reward directly from social signals in real-time during the interaction, to quickly learn about and dynamically address individual human preferences. Our scenario involves a Reeti robot in the role of a story teller talking about the main characters in the novel "Alice's Adventures in Wonderland" by generating descriptions with varying degree of introversion/extraversion. After initial simulation results, an interactive prototype is presented which allows to explore the learning process adapting to the human interaction partner's engagement.**

I. Introduction

With Socially Interactive Robots becoming more and more important in many research areas such as education, health and elderly care or in the context of domestic companions, creating customized and individualized interaction tailored to the human is one important challenge to increase robot acceptance. In the long run, a relationship between user and robot should be established, which is more likely to be achieved if the robot is equipped with a compelling personality, as this makes interaction more interesting and desirable [7]. However, there are different findings regarding personality preferences: research done by Aly and Tapus [1] indicates that adaptation of the robot's to the human's personality profile makes interaction more engaging and that humans prefer robots with similar personality [6], [5] to their own. In contrast, results by Joosse et al. [12] indicate that the task context plays a key role for giving an answer on whether a similar or opposing personality is preferred. Thus, it is not easy to tell which personality a robot should express in a given interaction scenario as this may depend on several factors, including the task itself and the user's preferences, which may also change over time.

To address this problem and to equip the robot with a personality tailored to the individual user's needs, this work proposes an autonomous learning process to find and adjust the optimal robot personality automatically, as, in general, adaptation is required to keep interaction engaging [27]. Several requirements have been taken into account: the process needs to run in real-time without any additional interaction as well as be unobtrusive to not disrupt the actual interaction, it should work in many application scenarios and also adapt accordingly if the human's preferences change.

While humans do not switch their personality in a very short time, they are able to a certain extent to portray a particular personality trait if required. For example, introvert persons may express enthusiasm in a similar way as extravert persons in order to advance their projects [17]. In a similar manner, the robot should be able to adapt its behavioral style if the situation calls for it.

A key challenge when implementing adaptation in Human-Robot Interaction (HRI) is to get feedback from the user to evaluate the robot's behavior. This is done in human-human interaction subliminally all the time by interpreting and reacting to gestures, mimics and other social signals expressed by the dialog partner. For a robot, it is much more complicated to get significant information suitable as an indicator on whether its behavior is pleasing, interesting, supportive or engaging for the user. Thus, many adaptation approaches rely on task-related information like the user's performance in a learning scenario to customize the robot's behavior. However, in some applications, there is no such relatively easy measurable data, especially when no explicit interaction goal exists, e.g. when the user does not have to solve a task like learning vocabularies or performing an exercise. As an example, the presented application at hand involves a Reeti robot which acts as story teller talking about characters in the novel "Alice in Wonderland" while the human is primarily listening.

To solve this problem, our adaptation approach combines traditional Reinforcement Learning (RL) [25] with subliminal feedback from the human: social signals. RL is used for real-time learning about the optimal robot personality and adapting to the desired profile over time instead of asking the user explicitly, sticking to a fixed personality or relying on mirroring the (opposite) user profile. The reward signal required for RL comes directly from the user's engagement, which is estimated based on multimodal social signal data.

This paper includes simulation results as well as an interactive human-robot dialog prototype to explore the adaptation process. A Reeti robot acts as story teller talking about characters in the novel "Alice in Wonderland". The robot's personality is expressed via linguistic style by generating utterances with varying amount of extraversion in real-time using Natural Language Generation (NLG). Based on the human's social signals, the online learning process controls and optimizes the robot's personality to keep the user engaged.

In section II, we discuss related work covering the com-

H. Ritschel, T. Baur and E. André are with Human-Centered Multimedia, Augsburg University, 86159 Augsburg, Germany {ritschel|baur|andre}@hcm-lab.de

978-1-5386-3518-6/17/$31.00 ©2017 IEEE

bination of social signals or rather personality and RL for adaptation in the context of HRI. Section III explains how the learning process works and how the RL task is modeled, as well as initial simulation results. Subsequently, section IV presents details on the prototype including sensing user engagement and NLG.

II. Related Work

In HRI, adaptation[1] may tackle different goals and address varying aspects of an interaction, for example learning long-term adaptation with focus on empathic supportive strategies [16] or affective behavior of a tutoring robot [11], adapting personality in the assistive domain for post-stroke users [28] or children with autism [18], where the system also has to be able to deal with the user's disabilities. While RL is often used as a learning framework, research that incorporates human social signals in the learning process, e.g. as a reward signal, is scarce. In this section, related work using RL in combination with social signals as well as adaptation of personality is discussed.

Liu et al. [18] use RL in the context of a basketball game for children with autism spectrum disorder. They adapt the robot's behavior to the child's affective state. To this end, the predicted liking level of the robot's behavior, which is estimated based on physiological signals, is used as a reward function. QV-Learning [31] was selected for RL.

Ferreira and Lefèvre [9] promote the concept of socially inspired rewards for online RL in a robot Dialog Management scenario. They suggest the use of human behavior cues as an additional reward signal at each dialog turn to speed up the policy optimization process and to enable adaptation to different users. This shaping reward, which is based on user appraisal, is added to the environment reward. In consequence, the reward signal is a combination of task-related data and human social signals. Throughout their experiments, the authors do not use real social signals, but a five-star rating bar on the user interface as workaround. Kalman temporal differences [10] is used for RL.

Leite et al. [16] use RL in the context of a chess companion for children which adapts its behavior to maximize the child's positive valence. The iCat robot learns the most effective empathic response depending on the child's affective state, which is determined by considering affective cues (smile, gaze) and task-related features (game evolution, chess board configuration from the child's perspective). This data allows to estimate the probability of positive feeling after and before employing different supportive strategies. The difference of those probabilities serves as reward for learning. Similar to Ferreira et al. [9], reward is a combination of task-related information as well as human social signals. The learning algorithm is based on n-armed bandit problems [25].

[1] Adaptation does not necessarily involve learning: there is research using robots in combination with social signals and/or biosignal information (such as in [26]) where the robot reacts (e.g. with a supportive reaction) to human behavior (e.g. a drop in user engagement) but does not learn its own behavior (e.g. which of the possible reactions is the best for the particular interaction partner). For the scope of this paper, we focus on work incorporating both adaptation *and* learning for personalization purposes.

While the research discussed above defines rewards as a combination of perceptible behavioral cues and task-related information, there are also RL approaches that compute rewards on the basis of perceptible behavioral cues alone. The later approaches are in particular suitable for applications in which humans communicate with robots just for the sake of interaction without a specific task in mind.

Gordon et al. [11] use RL in a student tutoring scenario with a smartphone and the Tega robot to maximize long-term learning gains. They estimate a child's valence and engagement based on facial expressions. Similar to Leite et al. [16], Gordon et al. do not use posture or gesture information. The state space includes the affective state of the child (discretized valence and engagement) as well as task-related information (whether the child interacted within the last 5 seconds and whether the last response was correct), while reward is based on child's valence and engagement only. By learning the robot's verbal and non-verbal behavior, the affective policy of the robot is personalized to the child. Traditional SARSA algorithm [23] is used for RL.

Barraquand and Crowley [3] learn appropriate behavior (politeness) depending on the current social situation. They rely on tactile feedback for reward and punishment, which can be triggered by caressing or tapping the AIBO robot's head or back. Classical Q-Learning [25], as well as multiple modifications, are used as learning algorithms.

Kim and Scassellati [13] use prosodic feedback to teach their Nico robot social waving behavior. The estimated prosodic affect is used as reward signal (approving or disapproving). Q-Learning is used to learn the optimal amplitude and frequency during waving of the elbow.

Mitsunaga et al. [22] realize adaptation of the Robovie-II robot's behavior to individual preferences in terms of interaction distance, gaze meeting, motion speed and timing. Their reward function is based on small discomfort signals from the human interaction partner, which are measured by the human's amount of movement and time spent gazing at the robot. Overall goal of the learning process is to minimize the user's discomfort. Mitsunaga et al. use the Policy Gradient Reinforcement Learning (PGRL) [15] algorithm.

Tapus et al. [28] use RL for behavior adaptation of an assistive therapist robot to the human's preferences. The robot's personality is expressed in terms of extraversion/introversion through vocal content and para-verbal cues. The authors use PGRL to manipulate movement speed, interaction distance and therapy style of the robot. Reward is calculated depending on the exercises performed throughout a time span, they do not take social signals into account.

To our best knowledge, the only work using RL to *learn* personality of a social robot is the one from Tapus et al. [28], where reward is calculated based on task-related information. However, the natural language utterances matching a particular personality style of the user have not been automatically generated in real time. We are not aware of any work that dynamically adapts linguistic style of a robot's automatically generated spoken utterances solely based on social signals that are interpreted as a reward.

Our contribution explores and implements the concept of reward based on social signals for real-time adaptation and personalization in application scenarios with hardly any measurable task-related information suitable as reward. In our prototype, we rely on user engagement, which is estimated by multimodal input including gesture and posture.

III. Adaptation Process and Simulation

There are five main requirements for our adaptation process: 1. real-time (run in parallel to the interaction, no waiting until the end to fill out a questionnaire, etc.) 2. no additional user interaction (e.g. rating current user experience) 3. unobtrusiveness (human user should not notice the existence of an adaptation process at all) 4. task-independence (should not rely on information dependent on the task) 5. reaction to changes in human's preferences over time.

We address the last requirement by using RL as a machine learning framework (see below) and the first four by making use of social signals for our adaptation approach. This data occurs in real-time during interaction anyways, which allows for adaptation without any extra effort which would interrupt the current task. Another important fact is that social signals are independent of the task. In their studies, Joosse et al. found out that it may depend on the task context whether a similar or opposing personality is preferred [12]. Since we do not know which personality profile is appropriate in a new application scenario without more ado, we do not stick to one personality but let the robot find out on its own. Our learning approach tries to maximize user engagement, which is applicable to many human-robot scenarios.

Inspired by the ARIA-VALUSPA project [29], we developed a dialog scenario with a Reeti robot as interaction partner that is able to hold a prolonged dialog with the user about the novel "Alice in Wonderland" by Lewis Caroll. Figure 1 illustrates the application: the robot presents facts about the main characters to the human user. Its personality is manipulated during interaction to find out whether the user prefers an introvert or extravert robot. For this purpose, descriptions are not predefined but generated by a NLG module at runtime: facts from the book context are transformed into utterances with varying amount of extraversion, one of the "Big Five" [20] personality dimensions. In parallel, social signals from the user are captured and analyzed with the Social Signals Interpretation (SSI) framework, which estimates current user engagement based on gesture, posture and video information. Changes in user engagement serve as reward signal for a RL process which optimizes the robot's personality to keep the user engaged.

RL is our algorithmic method of choice for learning about and adapting to the human user's personality preferences. This is due to two facts: first, RL is able to react to and learn about changes in human's preferences. Second, when relying on social signals, we have to handle different kinds of noise. The sensing hardware itself is subject to physical restrictions which limit the signals which can be perceived. Moreover, the interpretation process of sensed data can only be an approximation of the real user's engagement. Finally, the human's reaction to the personality expressed by the robot may vary from time to time as preferences may change, too. By using RL, which is an unsupervised learning approach, the robot is able to cope with this kinds of uncertainty. It is independent of instructive feedback since it explores the scenario autonomously based on trial and error.

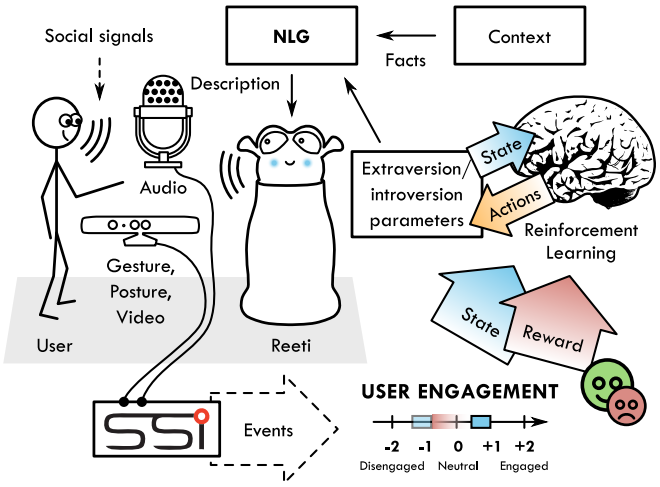

Fig. 1: Adaptation process.

A. Problem definition

From the perspective of the RL robot, goal of the adaptation approach is to maximize user engagement. User engagement at time t is defined as E_t. $E_t > 0$ indicates that the user is engaged, $E_t < 0$ that he or she is not happy with the interaction. Another important measurement is the change of user engagement between two sequent points in time $t-1$ and t, which is defined as $\Delta E_t = E_t - E_{t-1}$. This value indicates by which the amount of engagement changed during presentation of the last description. We expect an increase in engagement, i.e. $\Delta E_t > 0$, when the description was generated with a level of extraversion that is close to the user's actual preferences, and a decrease otherwise.

The robot's extraversion is defined as X, a value discretized in the integer interval $[-2; +2]$, which can be interpreted as *very introvert*, *introvert*, *neutral*, *extravert* and *very extravert*. X influences NLG parameters which cause the robot's utterances to be generated more introvert or extravert.

1) State space: The state space is kept as small as possible to realize quick exploration and adaptation. It is build up by two dimensions: estimated user engagement as well as the robot's current extraversion X, both integer values in the interval $[-2; +2]$. This allows the adaptation process to learn about the relationship between the robot's expressed personality and the user's engagement.

2) Action space: Since we concentrate on the extraversion/introversion dimension of the robot's personality, there are three actions to control the robot's personality: X can be increased or decreased by one or remain untouched. This limitation has two important advantages: 1. it prevents the robot from changing X too fast, which would cause generation of very diverging descriptions in terms of linguistic style

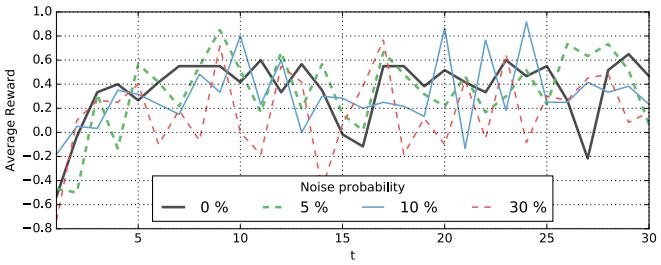

Fig. 2: Initial Q-Learning results.

within a very short time 2. it allows for faster learning since there are only three instead of five actions (when extraversion could be set to a specific value).

3) Reward: The reward signal for the adaptation approach is derived directly from social signals. It is defined as the change in user engagement ΔE_t mentioned above. If ΔE_t is greater than zero, engagement increased during presentation of the last description. Values smaller than zero indicate that engagement decreased, which means that the robot gets punished. The user's choice when accepting or rejecting a character suggestion does not influence the learning process at all, reward is based on user engagement exclusively.

B. Experiments

We conducted initial experiments based on Q-Learning with ϵ-greedy exploration to simulate the learning process. We use exploration rate $\epsilon = 0.2$ (high enough for handling noise) and learning rate $\alpha = 0.5$ (low enough to not eliminate all previous knowledge in case of noise). The simulated user's engagement increases when the robot's personality matches the actual preferences and decreases otherwise. Since such a deterministic user behavior is far from being realistic, noise simulates random changes in user engagement as well as deviations of the sensed from the real E_t value in each learning step. For $\Delta E_t = 0$, the robot gets a small reward +0.5 for preventing a decrease of engagement.

Figure 2 plots the averaged reward for every simulation step over 30 trials. Each trial corresponds to one unique user interacting with the robot over 30 time steps. In each trial, the robot presents 30 descriptions (30 learning steps) to a simulated user. At the beginning of each trial, the robot starts with neutral extraversion ($X = 0$) and an empty Q-Learning table (non-episodic learning task). The simulated user's personality preferences are initialzed randomly for each trial. To evaluate how much time is needed to adapt to preference changes algorithmically, user preferences always change at time step 15 and 26 during the experiments to a random value. This represents a worst case scenario as we expect preference changes to occur gradually over time in real experiments. However, it shows the agent's ability to adapt to the new preferences after a temporary performance loss. The initial random seed is the same for each noise probability experiment (the black line only averages over the 30 trials with no noise at all, the blue one over those with 10 percent probability, etc.).

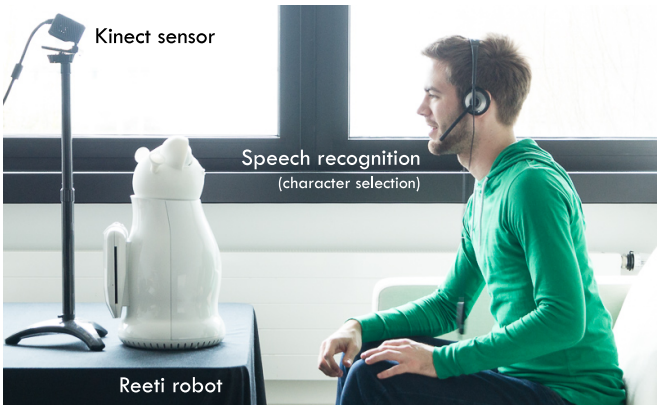

Fig. 3: Prototype scenario.

Without noise, learning is obviously quite robust. The average reward approaches 0.5: when the robot's extraversion level X equals to the user's preference, it learns not to change X anymore. Negative rewards can be attributed to exploration. Increasing noise leads to negative rewards occuring more frequently. Even if the graph looks noisy at first glance, it is important to notice that the averaged reward rarely falls below zero and is positive nearly all the time as long as there is not too much noise.

IV. Prototype

We built a live interaction prototype for the storytelling application to test the learning approach with real users. Both human user and robot sit opposite to each other (see figure 3). In the beginning, when the user greets the robot, it presents itself and suggests different characters to talk about. The user accepts or declines the character by speech commands and listens to the presented descriptions (see section IV-C). As soon as there is no new information left about a character, the robot suggests another character to continue.

During interaction, the user is captured with a Microsoft Kinect 2 sensor to process the user's social signals with the SSI framework (see section IV-B). RL is identical to the simulation with one exception: reward is calculated based on the non-discretized floating point difference ΔE_t, resulting in a more precise reward based on real social signals.

A. Dialog Scenario

Figure 4 illustrates the interaction between human user and robot. Reeti suggests a character (e.g. "Shall I tell you something about the white rabbit?") and waits for a response. This process is repeated until the user accepts a character. Then, from the perspective of the adaptation process, a new time step t begins. User engagement E_{t-1} at the time just before talking about the character is stored and the extraversion X is set by the RL process ($X = 0$ for $t = 0$). As soon as the NLG module is finished with generating a description according to X, the robot becomes active and presents the description. In the meantime, social signals are continuously interpreted to estimate user engagement. When the robot stops speaking, new user engagement E_t is used to

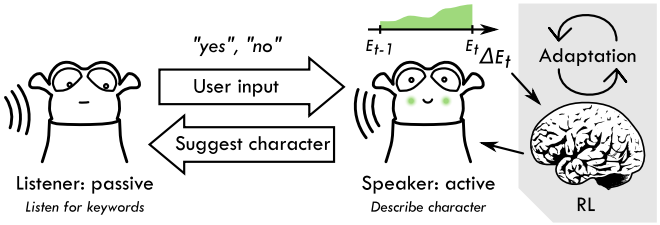

Fig. 4: Main dialog states.

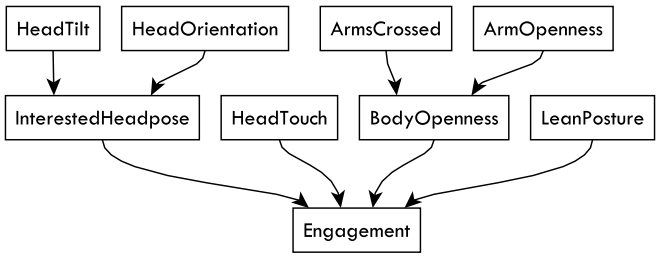

Fig. 5: Simplified Bayesian Network for user engagement.

calculate the reward $\Delta E_t = E_t - E_{t-1}$. Afterwards, the next time step $t+1$ begins. Descriptions for the selected character are generated as long as there is new information about it. Otherwise, the robot asks again in which of the remaining characters the user is interested.

In the context of RL, one time step corresponds to the presentation of one description, which takes several seconds depending on the utterance length. Therefore, the learning problem had to be formulated in such a way that it is possible to learn quickly and with limited experience to guarantee an acceptable user experience. Moreover, the robot's personality cannot deviate too much from the one in the last time step as extraversion can increase or decrease only by one. A change from maximum introvert to maximum extravert would require at least four time steps.

B. Estimating user engagement

For HRI, social signals play a key role. Sensing and processing the user's social cues is an essential part of the presented adaptation process and prototype. A Microsoft Kinect 2 sensor is used to capture the human user's movements in combination with the SSI framework [30] which processes and interprets the data in real-time. To calculate ΔE_t as reward signal for RL, we interpret the user's behaviors and derive the current engagement value E_t.

As suggested in [4] user's engagement E_t is estimated based on a Dynamic Bayesian Network (BN), a directed, acyclic graph with nodes representing variables and edges describing conditional probabilities [24]. Moreover, in Dynamic BNs temporal dependencies between the current state of variables and their earlier states can be modelled. Figure 5 shows an abstraction of a simplified BN for *Engagement*.

Each of the observed nodes contains two discrete states, *Present* and *Absent*. Based on the oberservations in these nodes, the probability of the values *Present* and *Absent* for the final node *Engagement* can be inferred. The network further contains "hidden" values that may not be directly

Fig. 6: An engaged and disengaged user interacting with the Reeti robot.

observed, but have to be inferred from observable variables. As an example, the likelihood that the variable *Interested Headpose* has the value *Present* is high if the value for the variable *Head Tilt* tends towards *Present* and the value for the variable *Look Away* would be close to *Absent*. The evidences for these values are constantly updated in real-time. For example, the probability that the variable *Arms Crossed* has the value *Present* is high if the corresponding social cue has been recognized with high confidence.

Cues that are considered relevant in our scenario are for example head tilt and orientation, which indicate whether the user is interested in the current interaction. The openness of the body is determined by the arm posture (opened or closed/crossed). E_t increases or decreases depending on gesture and posture. Figure 6 shows a user applying engaged and disengaged nonverbal behavior towards the robot. A user who leans himself forward is interpreted as more engaged as when he or she is leaned back. Further the amount of conversational regulators [8], such as back-channels indicate a high amount of engagement.

The BNs used in our system have been modeled with GeNIe[2]. The probablities of the variables in the network were learned based on the NoXi[3] corpus, which includes interactions of experts and novices about a certain topic, including Audio, Video and Kinect 2 depth streams.

Based on the value calculated by the BN, which is sent every 200 ms from SSI to the RL component, we further use a moving average with a five seconds window to smooth the estimated value. This allows us to calculate the user's engagement E_t in real time for each point in time t.

C. Generating descriptions

The NLG module is responsible for producing natural utterances for the robot that reflect a particular linguistic style in order to maximize user engagement. Previous studies have shown that user engagement is strongly influenced by an agent's personality. In addition, studies by Woods et al. [32] revealed that the extraversion personality trait plays an important role when users evaluate a robot's personality and assess to what extent it matches their own personality.

[2]http://genie.sis.pitt.edu/
[3]http://noxi.aria-agent.eu

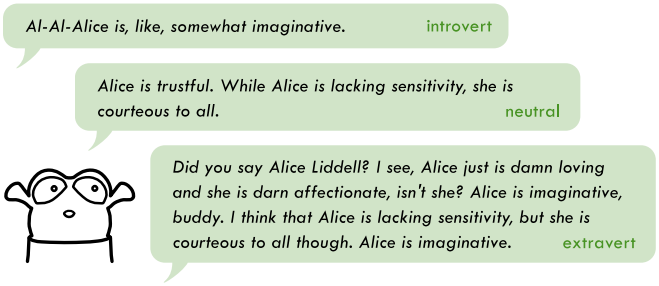

Fig. 7: Examples of generated descriptions.

Motivated by these studies, we decided to focus on extraversion as a personality trait and to adapt linguistic style to the amount of extraversion X, which is set by the RL module in each learning step. Information about the main characters is stored as facts and transformed into utterances in real-time. After presenting the generated description via the robot's Text-To-Speech (TTS) system, the RL approach manipulates X again and a new description is generated.

We use an approach inspired by PERSONAGE [19], a pipelined NLG system which realizes linguistic variation by applying different parameters in different stages of generation. Before generating a description, we set these parameters depending on X in order to generate utterances with a varying degree of extraversion (see below).

Figure 7 shows sample descriptions for maximum introvert, neutral and maximum extravert personality. One obvious difference between them is utterance length. This is mainly due to the content plan, which represents the basic structure of information. High extraversion implies high verbosity: a larger number of propositions (i.e. facts to present about the character) within one utterance is presented. Moreover, an introvert robot tends to use few positive emotion words and weakens positive content ("somewhat imaginative"), while an extravert robot employs many positive emotion words ("loving", "affectionate") and repeats itself ("imaginative"). Another characteristic of extraversion is the use of acknowledgments ("I see"), tag questions ("isn't she?") in-group markers ("buddy") and expletives ("damn"). Stuttering ("Al-Al-Alice") and using softener hedges ("somewhat") is typical for introvert utterances.

These are only a few personality-specific linguistic style features out of those we implemented. We refer to [19] for a complete set of parameters and details on how they map to extraversion or introversion. Finally, we use *SimpleNLG*[4] as surface realizer to convert the syntactical structure into a natural language utterance.

Parameters are set with some variation: the robot's extraversion X does not set or activate parameters directly, but increases or decreases their probability. Thus, we prevent the NLG module from using the same parameters for each description given a certain X, which would cause utterances to be stylistically too similar as long as the robot's personality does not change.

[4]https://github.com/simplenlg/simplenlg

V. Discussion

First experiments with the prototype enabled us to explore the time required for adaptation. The current knowledge base allows for an interaction time of roughly three minutes for describing Alice, the White Rabbit and the Queen of Hearts. During this time, the amount of learning steps depends on the robot's extraversion (an extravert robot presents more facts within one generated utterance). Thus, an introvert robot can learn quicker than an extravert because it will receive more rewards in the same time. When generating introvert (extravert) descriptions, roughly 30 (10) utterances are generated.

During this time, the robot learns very quickly because pronouncing the generated utterances only takes seconds. Whether three minutes are sufficient to explore the whole state space depends on the user and its expressivity in terms of showing and changing engagement. Since the learning algorithm only explores in 20 percent of its interactions and user engagement does not change extremely fast within this amount of time, we expect that more content is needed in real experiments to explore the complete state space. Moreover, the best point in time when to measure user engagement has to be investigated as changes may not occur immediately after the robot adjusts its amount of extraversion.

For our RL approach, we do not expect to converge to an optimal policy for every user. Several constellations are imaginable where convergence cannot happen. First of all, the human's actual preferences may change over time and with it the optimal policy, too. But especially the fact that RL is able to observe and react to these changes in its environment via exploration is essential for adaptation and explicitly desired. Moreover, depending on the application scenario, it may be counterproductive to excessively employ the same kind of linguistic style features since this could annoy the user, too. For example, an extravert robot using expletives could re-establish user engagement, but when using them too frequently, it may cause the opposite effect.

In scenarios with measurable information, a combination of both task-related and social feedback will probably be the most efficient and robust option of tuning a robot's behavior. However, there are robots that mainly serve to engage people in social interactions without pursuing a particular task-oriented goal, such as getting the user follow an advise. Furthermore, in some cases, it might take too long to wait until some achievements, such as progress in learning or rehabilitation, can be measured. That is the robot has to adapt its behavior dynamically while communicating with the user taking continuously provided user feedback into account. Finally, users might not be willing to take the effort to provide the robot permanently with explicit feedback. As Amershi et al. [2] pointed out users are not willing to "serve as an oracle" in order to teach the robot how to perform specific tasks. When engaging in a social dialog with a robot, providing oracle-like feedback to the robot might feel in particular unnatural. In such cases, falling back to social signal data seems to be a serious opportunity.

VI. Conclusion

In this paper, we proposed an approach based on RL and social signals for adapting the personality of a social robot to the preferences of the human user. While the robot acts as story teller talking about characters in the novel "Alice in Wonderland", its personality is manipulated in terms of extraversion expressed via NLG. The online learning process optimizes the robot's personality to keep the user engaged in the interaction. User engagement estimated based on multimodal social signal data serves as reward, which allows to use the approach in scenarios without measurable task-related information. Both simulation results and interactive prototype allow to explore the adaptation process.

In future work, we will investigate how to take advantage of implicitly and explicitly provided feedback in order to overcome the noise of continuously provided social signals and the sparsity of occasionally provided verbal feedback to the robot. By taking into account the user's social cues as a reward signal, the robot is able to adapt its behavior to maximize user engagement in a rather short time. However, as Knox and Stone [14] point out for task-based agents there is the danger of so-called "myoptic agents" that are not able to value human reward in the long run. To investigate this phenomenon for non-task-based dialog, we will move from single conversational episodes with the robot to multiple conversational episodes covering a longer period of time.

Acknowledgments

This research was funded by the Bavarian State Ministry for Education, Science and the Arts (STMWFK) as part of the ForGenderCare research association. Our work uses the WordNet® [21] database for NLG.

References

[1] A. Aly and A. Tapus. Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. *Autonomous Robots*, 40(2):193–209, 2015.

[2] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[3] R. Barraquand and J. L. Crowley. Learning polite behavior with situation models. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, HRI '08, pages 209–216. ACM, 2008.

[4] T. Baur, D. Schiller, and E. André. Modeling user's social attitude in a conversational system. In *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, Human–Computer Interaction Series, pages 181–199. Springer International Publishing, 2016.

[5] E. P. Bernier and B. Scassellati. The similarity-attraction effect in human-robot interaction. In *2010 IEEE 9th International Conference on Development and Learning*, pages 286–290, 2010.

[6] T. Bickmore, J. Cassell, T. Bickmore, and J. Cassell. Social dialogue with embodied conversational agents. In *Advances in Natural Multimodal Dialogue Systems*, volume 30 of *Text, Speech and Language Technology*, pages 23–54. Springer, 2005.

[7] C. Breazeal. *Designing sociable robots*. MIT press, 2004.

[8] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[9] E. Ferreira and F. Lefèvre. Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards. *Computer Speech & Language*, 34(1):256–274, 2015.

[10] M. Geist and O. Pietquin. Kalman temporal differences. *Journal of artificial intelligence research*, 39:483–532, 2010.

[11] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3951–3957. AAAI Press, 2016.

[12] M. Joosse, M. Lohse, J. G. Perez, and V. Evers. What you do is who you are: The role of task context in perceived social robot personality. In *2013 IEEE International Conference on Robotics and Automation*, pages 2134–2139. IEEE, 2013.

[13] E. S. Kim and B. Scassellati. Learning to refine behavior using prosodic feedback. In *2007 IEEE 6th International Conference on Development and Learning*, pages 205–210. IEEE, 2007.

[14] W. B. Knox and P. Stone. Learning non-myopically from human-generated reward. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 191–202. ACM, 2013.

[15] N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation, ICRA 2004, April 26 - May 1, 2004, New Orleans, LA, USA*, pages 2619–2624. IEEE, 2004.

[16] I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, and A. Paiva. Modelling empathy in social robotic companions. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 135–147. Springer, 2011.

[17] B. R. Little. Personal projects and free traits: Personality and motivation reconsidered. *Social and Personality Psychology Compass*, 2(3):1235–1254, 2008.

[18] C. Liu, K. Conn, N. Sarkar, and W. Stone. Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Transactions on Robotics*, 24(4):883–896, 2008.

[19] F. Mairesse and M. A. Walker. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488, 2011.

[20] R. R. McCrae and P. T. Costa. The five-factor theory of personality. In *Handbook of personality: Theory and research*, volume 3, pages 159–181. Guilford Press, 2008.

[21] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[22] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita. Adapting robot behavior for human–robot interaction. *IEEE Transactions on Robotics*, 24(4):911–916, 2008.

[23] S. P. Singh and R. S. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1):123–158, 1996.

[24] R. Stuart J. and N. Peter. *Artificial Intelligence: A modern approach*. Prentice Hall, 2003.

[25] R. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[26] D. Szafir and B. Mutlu. Pay attention! designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 11–20. ACM, 2012.

[27] A. Tapus, M. Mataric, and B. Scassellati. Socially assistive robotics [grand challenges of robotics]. *IEEE Robotics & Automation Magazine*, 14(1):35–42, 2007.

[28] A. Tapus, C. Tapus, and M. J. Matarić. User-robot personality matching and robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics*, 1(2):169–183, 2008.

[29] M. Valstar, T. Baur, A. Cafaro, A. Ghitulescu, B. Potard, J. Wagner, E. André, L. Durieu, M. Aylett, S. Dermouche, C. Pelachaud, E. Coutinho, B. Schuller, Y. Zhang, D. Heylen, M. Theune, and J. v. Waterschoot. Ask alice: An artificial retrieval of information agent. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 419–420. ACM, 2016.

[30] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (SSI) framework. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 831–834. ACM, 2013.

[31] M. A. Wiering. QV(λ)-learning: A new on-policy reinforcement learning algorithm. In *Proceedings of the 7th European Workshop on Reinforcement Learning*, volume 7, 2005.

[32] S. Woods, K. Dautenhahn, C. Kaouri, R. Boekhorst, and K. Lee Koay. Is this robot like me? links between human and robot personality traits. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pages 375–380. IEEE, 2005.