Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3164–3173
Marseille, 20-25 June 2022

ProDial – An Annotated Proactive Dialogue Act Corpus for Conversational Assistants using Crowdsourcing

Matthias Kraus, Nicolas Wagner, Wolfgang Minker
Department of Communications Engineering, Ulm University
Albert-Einstein-Allee 43, 89081 Ulm
{matthias.kraus, nicolas.wagner, wolfgang.minker}@uni-ulm.de

Abstract

Proactive behaviour is an integral interaction concept of both human-human as well as human-computer cooperation. However, modelling proactive systems and appropriate interaction strategies are still an open quest. In this work, a parameterised and annotated dialogue corpus has been created. The corpus is based on human interactions with an autonomous agent embedded in a serious game setting. For modelling proactive dialogue behaviour, the agent was capable of selecting from four different proactive actions (None, Notification, Suggestion, Intervention) in order to serve as the user's personal advisor in a sequential planning task. Data was collected online using crowdsourcing (308 participants) resulting in a total of 3696 system-user exchanges. Data was annotated with objective features as well as subjectively self-reported features for capturing the interplay between proactive behaviour and situational as well as user-dependent characteristics. The corpus is intended for building a user model for developing trustworthy proactive interaction strategies.

Keywords: assistance system, proactivity, dialogue corpora, dialogue model, human-computer trust

1. Introduction

Due to technological advancement and an ever growing market for digital assistants (Statista, 2016), it can be expected that conversational assistants (CA) enter highly sophisticated domains and be used for very challenging tasks, like decision-making (Peng et al., 2019), learning (Paladines and Ramírez, 2020), or planning tasks (Behnke et al., 2020). To be accepted and trusted in these delicate domains, conversational interfaces must extend their assistance capabilities and be equipped with more human-like assistant behaviour. An important feature of human assistants or advisors forms the concept of proactivity. Proactive behaviour, i.e to make assumptions about future situations and to act in advance accordingly instead of only reacting, is fundamental both in cooperative human-human as well as human-computer interaction (HCI) (Crant, 2000; Nothdurft et al., 2015). A key capability of proactive systems is the ability to foreshadow possible user behaviour. Considering a decision-making scenario, for example, this would imply that a proactive assistant would either make a selection in the user's interest or at least provide appropriate suggestions. Such system behaviour could be quite beneficial with regard to task efficiency, user satisfaction (Baraglia et al., 2016; Peng et al., 2019; Kraus et al., 2020a), and overall trust in the system (Rau et al., 2013; Kraus et al., 2020c), as opposed to a reactive assistant that would only help on explicit user request. On the other hand, the system's decision whether to become proactive and to which extent is required to be considered carefully as misuse could turn out badly. The arguably best negative example for this case is Microsoft's former office assistant Clippit. The early proactive assistant interrupted users at inappropriate moments during task execution, providing non-helpful assistance, while behaving highly obtrusive (Bickmore and Picard, 2005). This ultimately lead to distrust and the rejection of the system, and hence stresses the importance and complexity of the development of sound proactive interaction strategies. Thereby essential is the timing for initiating a specific kind of proactive dialogue in order to not harm the human-computer relationship. For this, the system has to weigh costs and benefits for decision-making. Here, rule-based approaches tend to fail, as the need for proactive behaviour is highly user- and situation-dependent (Nothdurft et al., 2015; Kraus et al., 2020c). Hence, data-based statistical methods are required to incorporate user and context knowledge for proactive dialogue modeling.

Although there exists a variety of data corpora (e.g. DSCT (Williams et al., 2014), MultiWoz (Budzianowski et al., 2018)) for conventional dialogue modeling, none of them are sufficient for modeling proactive dialogue as proactive behaviour is simply not included or highly under represented (Balaraman and Magnini, 2020). To counteract this gap in current literature, a corpus for developing proactive dialogue strategies is presented in this work.

For generating the corpus, data was collected online with 308 real users who had to interact with an artificial advisor agent in a serious proactive dialogue game. The user's task in this context was to take the role of a CEO and make decisions for successfully establishing a new company. For modeling proactive dialogue behaviour, an agent system was capable of selecting from four different proactive actios (None, Notification, Suggestion, Intervention) in order to serve as the user's personal advisor. The data was annotated with objective features, e.g, task duration and success, as well as subjectively self-reported features, e.g. user's age, gender, and personality. As trust forms a fundamental concept for the acceptance of autonomous agents and therein proactive dialogue systems, interactions in the corpus were labeled with self-reported measures on the system's trustworthiness and its related concepts competence, reliability, and predictability (Madsen and Gregor, 2000). In addition, users could self-report their satisfaction with the system's performance and its level of annoyance. In summary, this paper makes the following contributions: First, the development of a simulated environment for proactive system decision-making is described. Secondly, a novel method for collecting personal and dialogue data in a serious gaming scenario is presented. Finally, the main contribution is an annotated data corpus containing interactions with the proactive assistant system. The corpus is ought to be used for developing a user model for predicting the perceived trustworthiness of a proactive system. Amongst other, trust prediction could be used for developing user-adaptive proactive dialogue models through automatically learning a proactive dialogue policy via reinforcement learning by utilising trust as a reward.

2. Related Work

2.1. Methods for Collecting Dialogue Data

According to Budzianowski et al. (2018), there exist three types of methods for collecting dialogue data: machine-to-machine, human-to-human, and human-to-machine. Machine-to-machine dialogue data is collected by simulating interaction outlines between an artificial user and a system bot via dialogue self-play (Shah et al., 2018). For generating a more diverse data set, crowd workers are then recruited for paraphrasing the utterances. This approach is useful for generating data for building task-oriented dialogue systems, that is one of the primary research fields in dialogue research, where the focus is set on database-querying tasks and slot-filling dialogues (Young et al., 2013; McTear et al., 2016). However, procedural turn-taking, i.e. a sequential planning or decision-making task as utilised in this work, as well as the handling of unstructured data is not covered by this approach. This approach is also highly influenced by the quality and capabilities of the user simulator. As to this date there exists no user simulator handling proactive dialogue and which is able to simulate trust, this approach is not applicable for our work.

The arguably ideal way of collecting dialogue forms collecting data form human-human interactions, being the most natural approach and providing a high diversity of dialogues. With the rise of social networks, the idea of recording publicly accessible conversations emerged. Relying on unsupervised clustering algorithms, the Twitter data set (Ritter et al., 2010) consists of an open domain collection with more than a million of conversations extracted from Twitter. Similarly, the Ubuntu corpus (Lowe et al., 2015) provides chats in the area of technical support. The disadvantage of this type of collections is that parts of the data contain unusable texts and spellings. Additionally, the majority of these conversations are not goal-oriented dialogues, while being mostly used to train end-to-end dialogue systems (Lowe et al., 2017). A special type of human-to-human data collections are Wizard-of-Oz (WoZ) datasets (Kelley, 1984), in which a human wizard simulates system behaviour. Here, dialogues follow a pre-defined script designating the potential actions the wizard is allowed to take. The simulated system behaviour appears thus logically consistent and human-like. To increase the quality and diversity of dialogues, the MultiWOZ approach (Budzianowski et al., 2018) was developed containing 10k dialogues in different domains obtained by employing crowd workers. The procedure for collecting such a dataset is, however, cumbersome and resource-intensive, since it requires additional human work-time for the data collections as well as for the subsequent transcription and labelling. This considerably limits the total number of samples and can lead to an inconsistent system due to the dissimilar behaviour of different wizards.

As the last type, human-to-machine data collection is conducted with interactions between users and existing dialogue systems. Naturally, the prerequisite for this approach is that a dialogue system is available and that all necessary functions have been implemented. This in turn facilitates the annotation process as it is possible to extract objective features directly during the experiment. So far, there already exist quite a number of such developed corpora, including the "Let's Go Bus Information system" corpus from the Carnegie Mellon University (Black and Eskenazi, 2009). On the basis of this corpus, Schmitt et al. (2011) developed a prediction model for estimating the interaction quality of the dialogues. The underlying structure of the dialogues form user-system exchanges, i.e. a user utterance is followed by a system response or vice versa in a consecutive manner until one party finishes the dialogue. This format seems to be particularly suitable for sequential planning and decision-making tasks while interacting with an assistance system (Biundo and Wendemuth, 2016). Therefore, the decision-making task for the data collection is represented in the system-user exchange where an assistant is able to proactively influence a user's decision. Additionally, this allows annotations on an exchange basis enabling a structured labeling process of proactive system actions. More insight on proactive HCI in general is presented in the following section.

2.2. Proactive Human-Computer Interaction

In this work, we consider proactive dialogue in the sense of mixed-initiative collaboration between a conversational system and a user (Horvitz, 1999). In such collaborations, a user and an autonomous agent act as team for solving tasks, where each partner may take actions independently. In such scenarios, the agent needs to track the user's activities and goals while reasoning about the costs and benefits of taking automated actions. Here, proactive dialogue serves for communicating and negotiating a system's decision process in order to minimise the risk of system failure. For example, the proactive assistant CALO (Yorke-Smith et al., 2012) makes use of reasoning for estimating the cost-benefit value of proactive behaviour and adjusts its actions. The cost-benefit function is based on system-related metrics, e.g. time-sensitivity of the suggestion, the degree of uncertainty, and the system's confidence. CALO is able to choose of set of proactive actions that can be differentiated with regard to their level of intrusiveness. The lowest level represents reactive behaviour, i.e. system acts upon user request, whereas the highest level represents completely autonomous system behaviour, i.e. the system acts on behalf of the user. In between, the system can make suggestions. This approach of modelling proactive actions originates from the well-known levels of autonomy introduced in the seminal work by Sheridan and Verplank (1978) and is commonly used in human-robot interaction (Baraglia et al., 2016; Peng et al., 2019) and user interface design (Isbell and Pierce, 2005). In this work, we also used a set of levelled proactive dialogue actions.

While CALO's decision which level to choose is based on hand-crafted thresholds, we deem the problem of selecting an appropriate proactive strategy to greatly deal with decision-making under uncertainty. Hence, data-driven approaches could provide a more effective way for generating sound proactive behaviour. Therefore, a data corpus of mixed-initiative human-computer collaboration including different levels of proactive dialogue actions was generated in the scope of this work. The corpus is intended for creating a proactive dialogue model. An essential part of this model is an adequate representation of the state of the user. Kraus et al. (2020b) discovered that the user's perception of an proactive assistant differed depending on specific user characteristics, such as the technical affinity or previous experience with dialogue systems. Thus, the corpus was annotated with personal user information and self-reported user experience measures for quantifying the subjective effect of the CA's behaviour. Here, the main focus is set on the human-computer trust (HCT) relationship, which has shown to correlate with the degree of proactive behaviour of a technical system. For example, Rau et al. (2013) compared two levels of autonomy, high versus low. The authors presented a WoZ-study, in which participants had to complete a sea survival task in collaboration with a remotely controlled robot. The study results demonstrated that trust in the robot was higher in the low-level (reactive) than in the high-level condition. In another study (Kraus et al., 2020b), it was found that proactive dialogue showed strong effects on cognition-based or performance-based trust (Madsen and Gregor, 2000), i.e. system's perceived competence and reliability, depending on task difficulty. Similarly to the results presented in Rau et al. (2013), highly autonomous system behaviour did not foster a HCT relationship, in contrast to conservative actions. For a better understanding of the concept of trust, an overview of related work on HCT is presented in the following.

2.3. Human-Computer Trust

Trust is an important factor in HCI, where fundamental work has been provided in the field of autonomous systems. Trust in automation can be defined as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p.51). The definition suggests three fundamental factors for modelling trust: the human, the autonomous partner, and the environment (Schaefer et al., 2016). Personal user characteristics that influence trust form, for example, individual traits (e.g. gender, personality), states (e.g. stress, fatigue), cognitive factors (e.g. technical understanding, expectancy). A person's trust propensity can serve as a baseline for predicting the initial HCT level (Merritt and Ilgen, 2008; Jian et al., 2000). Factors of the autonomous partner that influence trust are system specific (e.g. level of automation, anthropomorphism) and capability related features (e.g. reliability, competence). Further, environmental factors such as the task/context (e.g. task difficulty, task type) need to be considered. As there exists various models for HCT, the model developed by Madsen and Gregor (2000) is used in this work. According to this model, trust is mainly formed by cognition (performance)- and affect-based trust. Affect-based trust comprises the user's personal attachment towards and faith in the system. As these concepts require some experience and knowledge about the system, affect-based trust refers to long-term relationships. Contrary, cognition-based trust refers to a more short-termed trust. Here, mostly the performance of a system are of importance. Hence, the bases for cognition-based trust comprise the user's perceived competence, reliability, and understandability/predictability of the system. In this paper, a focus is set on cognition- or performance-based trust due to the short-term interactions with the developed system. For assessing the trust, primarily subjective measurements in the form of self-reported questionnaires are collected (Madsen and Gregor, 2000; Gulati et al., 2019). The data corpus presented in this work was annotated by crowd workers with subjective self-report measures using a Likert scale. For not straining the users' cognitive loads and thus limit the risk of generating unreliable and problematic annotations, the labeling process was clearly separated from system interaction.

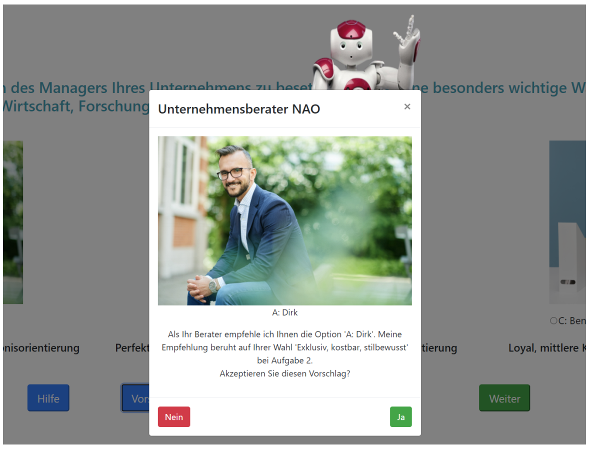
3. Data Collection Method

For the human-to-machine data collection, a prototypical proactive dialogue assistant was developed based on our previous work (Kraus et al., 2020c). The assistant was embedded in a mixed-initiative serious games environment. Serious games are “games used for purposes other than mere entertainment” (Susi et al., 2007), and are intended to lever “the power of computer games to captivate and engage end-users for a specific purpose, such as to develop new knowledge and skills” (Corti, 2006). Two properties of serious gaming are particularly beneficial for the approach of data acquisition presented in this work: First, serious games are highly motivating for users and foster engagement and intrinsic motivation (Abt, 1970). Engaged users are required in order for them to take the game and the assistant’s actions seriously. In doing so, an environment of risk and vulnerability was ought to be created. Thus, trust in the system could be developed or destroyed depending on the agent’s actions in such an environment. Secondly, actually testing and evaluating policies (or in our case dialogue strategies) in the real world is too expensive and cumbersome. For this reason, serious games provide a simulated reality based on reduced-scale models for allowing problem-solving (Abt, 1970). Hence, such games enable to evaluate the consequences of alternative dialogue policies on the HCT in different situations, promoting the development of data-driven adaptive strategies.

The data acquisition itself was conducted online with crowd workers, that interacted with the system and provided annotations regarding their perception of the system at defined time steps. Objective features were automatically collected by the system itself and combined with user annotations to be written to a database. In the following sections, a detailed description of the serious game scenario and an overview of the prototypical system is provided. The scenario and data collection has been previously outlined in an earlier work of ours (Kraus et al., 2021), however is explained more in detail in this work for providing a full description of the data collection process and the created corpus.

3.1. CEO – A Serious Proactive Dialogue Game

A role-playing game was selected as a scenario, in which a user took the role of the CEO of a high-tech company that develops, produces, and sells electrically powered cars. The user’s goal was to successfully manage the company by executing strategic actions for maximising profits. In doing so, users had to make step-by-step decisions and plan undertakings in the interest of the company, such as location planning or personnel management. Individual decisions had consequences and affected the success of the management. The game was designed as a turn-based planning task, in which the system sequentially presented a task step and the available choices, whereas the user could take different actions and cooperatively solve the task with a CA. Hence, the structure of the game resembles that of a system-directed dialogue in which both dialogue participants take turns, i.e. the system takes an action providing task step relevant information, upon the user takes an action solving the respective task step. The game ended after a total of 12 task steps. The order of the tasks was fixed and could not be altered by the user. For each step, several options from which the user had to select were presented. The number of options changed from task to task ranging from a minimum of three to a maximum of five options. The purpose for this was to vary the complexity of each task in order to influence the user’s perceived task difficulty, as this has shown to effect the perceived trustworthiness of proactive behaviour (Kraus et al., 2020c). At each task step, users could execute four actions: select an option without system assistance and continue with the game, explicitly ask the CA for a suggestion, or ask for help. By asking for help, general information about the game is provided, e.g. which previous decisions need to be considered at the current task step. By asking the assistant for a suggestion, appropriate advise was provided. The user could either accept or decline the system’s proposal. When an answer option had been selected, the user could continue with the game. The success of the user’s decision-making was measured by attaching numeric scores to the individual selection options. This allowed previous decisions to directly influence the value of future actions. Consider the following example: User Alice is currently required to make a decision on task “Research”, where a plausible research direction with regard to the built up company needs to be chosen. This task is influenced by Alice selections in previous tasks “Management” and “Banking”. Depending on the combination of selections in respective tasks, one of the four options (Hydrogen Drive, Autonomous Driving, Battery Research, Climate Neutral Production) would yield the most points, whereas in the worst case scenario a user would yield zero points


Unternehmensberater NAO

A: Dirk

Als Ihr Berater empfehle ich Ihnen die Option 'A: Dirk'. Meine Empfehlung beruht auf Ihrer Wahl 'Exklusiv, kostbar, stilbewusst' bei Aufgabe 2.
Akzeptieren Sie diesen Vorschlag?

Figure 1: Illustration of the proactive assistant and its suggestion during the CEO-game (Kraus et al., 2021).

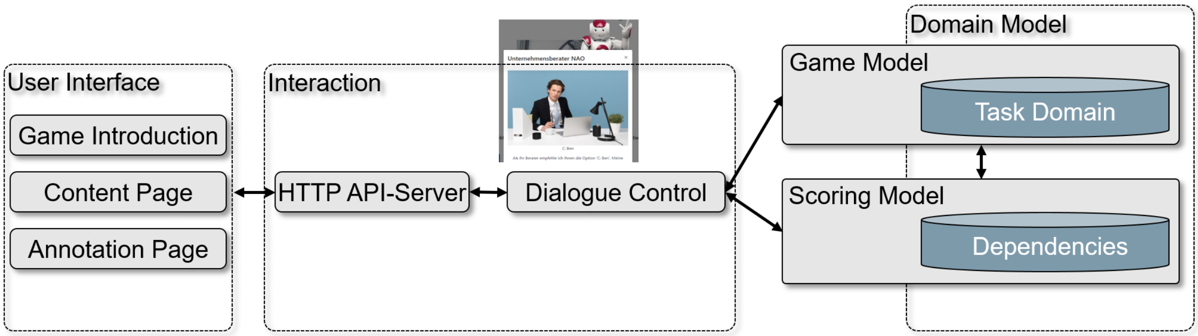


Figure 2: Overview of the system's architecture for creating a data collection environment.

minimum. The concept of a game score should create a vulnerable but also engaging environment for the user. Thus, performance was used as incentive for the users to take the game seriously. The game score was based on an artificial scoring model, particularly developed for this application.

During the game the user was supported with decision-making by a CA. The agent was introduced to the user as an AI-driven virtual assistant called NAO, who provides business advice. An anthropomorphised assistant was chosen in order to form a clearer separation of task and assistance technology compared to using simple pop-up messages. For this, the assistant was presented as a picture of the famous humanoid robot NAO from Softbank Robotics. NAO could either provide suggestions in an active or reactive fashion. A depiction of the CA and its proactive text messages is presented in Fig. 1. NAO was designed to be an expert system avoiding the unintended side effects of incompetent system behaviour on its trustworthiness. This allowed to only consider the effects of the proactive levels on the HCT. For selecting the best option per task step, the assistant made use of a simple reasoning mechanism by having knowledge about past user selections and accordingly querying the game's scoring model. Further, proactive explanations have been added to justify the behaviour of the system to take the initiative. For creating the explaining messages, a template-based approach was used. The relation between the best option and previous user selections that led to finding this option was exploited to include information about past user behaviour in the explanation. This information was transformed into natural language and wrapped into predefined sentence template. For example, "As your adviser, I recommend option A. My recommendation is based on your choice of B in Task C, whose characteristics of D best fit our concept".

In the context of mixed-initiative interaction, proactive behaviour implicates that the CA suggests or takes over action's on behalf of the user. Therefore, proactive actions can be considered as the initiation of sub-dialogues, where the assistant influences a user action.

3.2. System Design

The CEO-game was implemented as servlet-application based on a client-server model. On the client-side a user played the game and interacted with the proactive assistant using a clickable graphical user interface (GUI). On the server-side a dialogue control logic received user input from the GUI and provided task-related content to the interface by accessing a database. The JSON-based database served as model for the application and contained the complete game-content and structure as well as the scoring model. Information between GUI and dialogue control was exchanged using HTTP client requests and Javascript forms. The system's architecture is visualised in Fig. 2. The GUI was created as an HTML/Javascript-based web page. The web page's content was created dynamically on the fly for each task retrieving content from the database through the dialog control. In general, the tasks were presented on the GUI using title and number of the current task, a task description, and the different task options (name, image). User's could interact with the GUI using its action buttons (help, suggest, select option, continue). The action buttons were blocked for 20 seconds for providing the user with enough time to read the information about the task and to guarantee that the user received the system's proactive messages which were triggered after the same amount of time. In this way, the CA did not interfere with the user's process of getting familiar with the task description.

The dialogue control logic was implemented as an HTTP web server. It was responsible for controlling the (proactive) interaction with the user and stored information of a game session, while interacting with a game-specific database for retrieving relevant domain information. The database contained models for the game-content and structure and scoring model, both defined in the JSON format. These models were intended to simulate the necessary artificial intelligence components for allowing proactive behaviour – planning (game model) and reasoning (scoring model) (Behnke et al., 2020). The game model consisted of a sequence of task steps comprising the

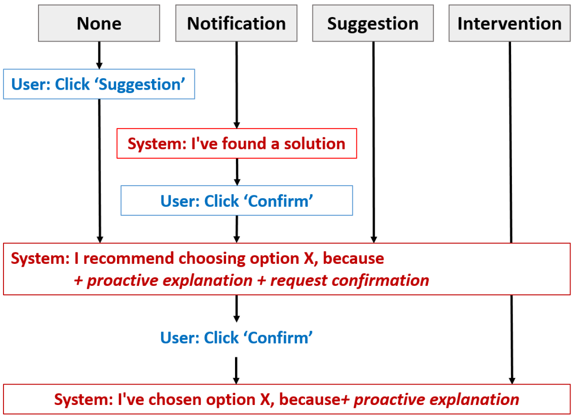

None
Notification
Suggestion
Intervention
User: Click 'Suggestion'
System: I've found a solution
User: Click 'Confirm'
System: I recommend choosing option X, because
+ proactive explanation + request confirmation
User: Click 'Confirm'
System: I've chosen option X, because+ proactive explanation

Figure 3: Flowchart, visualising the dialogue content of different levels of proactivity. User utterances are coloured in blue, while system actions are red-coloured (Kraus et al., 2021).

different options for each task step and associated information.

In correspondence to the game model, the scoring model was also constructed of a sequence of task steps. However, the individual task steps comprised information about the influence of previous decisions on the options of the current task step. This concept was called "Dependencies". If one of the options is positively influenced by a previous selection it is valued with a score of 10, otherwise it receives a score of 0. This composition allowed the dialogue control to determine the best selection of the current task step with regard to previous choices. The CA then exploited this knowledge when providing reactive or proactive assistance. Proactive assistance was modelled based on the proactive dialogue actions defined in our previous work (Kraus et al., 2020c): *None*, *Notification*, *Suggestion*, *Intervention*. The individual actions range from no automation to complete automation. The purpose of the assistant was to provide helpful guidance in the selection process. Using the reactive *None*-action, the system awaited user requests for making suggestions. The more conservative proactive actions *Notification* and *Suggestion* let the user confirm the assistant's proposals and differ solely in their level of directness. The *Intervention*-action took over responsibility and autonomously selected an option. The dialogue flows of the different actions are depicted in Fig. 3.

For gathering a sufficient amount of data, the proactive dialogues were initiated at random task steps using a restricted randomising policy for ensuring naturalness. The policy restricted proactivity to occur only on four out of twelve task steps, as too frequent system interventions were deemed unrealistic and annoying in a counseling scenario. Further, proactivity was restricted to occur only once within every three task steps for facilitating the annotation process. The annotation process is detailed in the next section.

3.3. Annotation Process

The annotation process was divided into two phases. First, users provided anonymised personal information by answering a questionnaire. Secondly, for collecting data on the perception of the proactive dialogue assistant, users were instructed to rate their experience with the system, e.g. trust, competence, user satisfaction, after every three tasks. By not measuring trust at each step, it was supposed to prevent survey fatigue of the users and to preserve the participants' cognitive loads. The CA was designed to become proactive at one task step during each segment of three tasks. This was intended to capture the effect of one specific proactive level at a time. The specific task step and specific task step was selected randomly by the system using a uniform distribution. During the decision-making users were unaware of the consequences of the respective option selection, nor did they know about the expertise of the CA. This method was ought to ensure the vulnerability of the user towards the assistant and let the user self-explore the abilities and usefulness of the system. For helping the users to rate the experience with the assistant, the outcome of their choices were presented in the form of a game score after every three steps. The user experience ratings were then attached to the previous three exchanges. For obtaining a sufficiently large distribution of different proactive actions across all steps, a high number of user interactions was required. Objective annotations, like task success, duration, or the user's actions taken, were captured by the system itself at each task step.

4. Corpus Information

A corpus for the creation of user-adaptive proactive dialogue was generated by collecting data from user interactions with the described conversational agent during the CEO-game. The data collection was centred around features known to be effecting the user's trust in the CA. Characteristic features from previous work on trust in automation and HCI in general were selected. Based on literature research, 10 user-dependent and 7 context-dependent features were considered for predicting the user's trust. An overview of the features is described in Table 1. User-dependent features were collected using a questionnaire before the user had started the game and therefore remained static throughout playing the game. In the questionnaire, users could state their age by providing a numeric value. Three options (female, male, diverse) were possible for expressing the gender. Personality information was collected using the BFI-10 scale developed by Rammstedt et al. (Rammstedt et al., 2013). This scale measured the extent of the personality traits of the well-known Big-Five inventory neuroticism, extraversion, openness, agreeableness, and conscientiousness (McCrae and John, 1992) scale. Affinity towards technical systems was assessed using the TA-EG-scale comprising six statements designed by Karrer et al (Kar-

User-Dependent Features	*Context-Dependent Features*	*Target Features*
Age	Proactive Action	Trust
Gender	Task Difficulty	Competence
Technical Affinity (Karrer et al., 2009)	Task Complexity	Reliability
Trust Propensity (Merritt et al., 2013)	Task Duration	Predictability
Domain Expertise	User Selection	Acceptance
Big 5 personality traits (Rammstedt et al., 2013)	Suggestion Request	Annoyance
	Help Request	User Satisfaction

Table 1: Overview of the collected features using the described data collection method.

rer et al., 2009). For measuring domain expertise, we developed our own questionnaire consisting of three items for checking the user's experience in management. Further, propensity towards trust in autonomous systems with the scale by Merrit et al. (Merritt et al., 2013) was measured in order to gain information about the user's initial trust. All scales were measured on 5-point Likert scales. Context-dependent features were collected for each of the twelve task steps. Proactive actions were annotated at each step in the format None, Notification, Suggestion, or Intervention. Perceived task difficulty was self-reported by the user on segment-level, i.e. after three task steps, using a 5-point Likert scale ranging form 1="very low" to 5="very high". Task complexity denotes the amount of options of a specific task step and ranged from three to five. User selection indicates the amount of points a user received for his or her decision at a task step. The minimum points a user can receive is zero, while it possible to gather a maximum of 40 points for one decision. Task duration was measured in seconds for each task step. When the user triggered a suggestion or a help request for a certain task step, either a 1 (= action triggered) was annotated. Otherwise, a 0 (=action not triggered) was noted. The target variable trust was measured on a 5-point Likert scale after each segment of the game for the reasons described in the previous section. The scale ranged ranging 1="very low" to 5="very high". The annotated trust value was also applied to the previous three task steps. We deemed trust to stay invariant during this time frame as only one proactive action was triggered. Additionally, the user's perceived competence, predictability, reliability to represent the user's cognition-based trust (Madsen and Gregor, 2000) were annotated. Data collection was conducted using the German clickworker platform [1]. Eligibility conditions required users to be aged between 18 and 60, to be a native speaker of German, and to play the game on a desktop computer for compatibility reasons. In total 320 participants were recruited. However, twelve had to be excluded due to violation of instructional terms and technical errors resulting in a final number of 308 users for data collection. In advance of the start of the game, user's were briefed about details of the data survey, e.g. duration (20 minutes) and purpose of the survey. Further, participants were informed that concentration checks were included in the ratings to take the game and the evaluation seriously. When users did not pass the checks they did not receive their reward. Participation was compensated with a monetary reward of 3 €. Further, the actions buttons were blocked for 20 seconds to avoid that users click through the tasks. The details of the corpus are depicted in Table 2. Overall, the agent was rated generally as trustworthy with 52 % of the exchanges with the

#Dialogues	308
#System-User Exchanges	3696
Avg. Dialogue Duration in seconds	492 s ± 191
Avg. Duration System-User Exchange in seconds	41 s ± 16
Avg. Perceived Task Difficulty	2.6 ± 0.6
Avg. #Help Clicks	0.6 ± 1.8
Avg. #Suggestion Clicks	5.2 ± 3.3
Avg. #Total Points	154 ± 28 / 210
#Proactive-None	2523
#Proactive-Notification	364
#Proactive-Suggestion	419
#Proactive-Intervention	390
Avg. User Age in years	37 y ± 11
#Male	194
#Female	113
#Other	1
Avg. Technical Affinity	4.0 ± 0.5
Avg. Experience Management	2.9 ± 1.0
Avg. Propensity to Trust	3.5 ± 0.7
#Trust-Very Low	69
#Trust-Low	336
#Trust-Neutral	1242
#Trust-High	1707
#Trust-Very High	342

Table 2: Descriptive statistics of the generated corpus. Counts are symbolised with the prefix #.

[1] www.clickworker.de

system were labeled with "High" or "Very High" trustworthiness. Consequently, the expert assistant system was able to provide adequate help as was expected per design. More interestingly, even though the assistant always provided a correct suggestion, still 11 % of the system exchanges were rated with below neutral trustworthiness. This could be explained either by inappropriate proactive system behaviour or by a user's general low tendency to trust a technical system irrespective of its capabilities. However, the tendency to use the agent for help is evident by considering the amount of suggestion clicks. Requests for the system's suggestion messages were used 43 % of the dialogue. Hence, this may be more related to the random dialogue strategy of the system. Requests for help regarding the principle of the game were used rarely (5 % per dialogue). This indicates a clear and understandable design of the developed dialogue game.

For evaluating the usefulness of the collected data, we investigated whether there exist the same correlations between user- and system-related factors in the corpus as found in related work. In line with related work, the user characteristics that correlated the most with trust were the user's propensity to trust a technical system ($Spearman's\ r = 0.32,\ p < .001$) and technical affinity ($r = 0.23,\ p < .001$), e.g. see Merritt et al. (2013) and Kraus (2020). While the user's age ($r = -.005,\ p = 0.76$) and gender ($F = 0.84,\ p = 0.36$) did not generally correlate with trust and its related concepts (also mentioned in Hoff and Bashir (2015)), the domain expertise of an individual user showed a significant relationship ($r = 0.13,\ p < .001$). In contrast to related work (Sanchez et al., 2014), where a higher domain expertise related to a lower trust in the technical system, a positive correlation was found in the corpus. This indicates that the content of the agent's suggestions were indeed implemented adequately. Further, significant correlations between the user's personality and the HCT were discovered. In line with related work, a positive correlation between extraversion (Evans and Revelle, 2008) ($r = 0.12,\ p < .001$), agreeableness ($r = 0.09,\ p < .001$), and conscientiousness ($r = 0.16,\ p < .001$) (Chien et al., 2016) and trust was found. Additionally, a negative correlation between neuroticism and trust was found ($r = -0.10,\ p < .001$). This was also indicated by a previous study of Evans and Revelle (2008).

The proactive actions did not differ significantly regarding their influence on the system's perceived trustworthiness ($F = 0.98,\ p = 0.40$). Consequently, there seems to exist no "one size fits all" solution on designing proactive dialogue strategies. However, this could be expected as we randomly triggered the proactive actions without taking into account user features or context. For designing proactive strategies, the personal and context information gathered in this work may be used for creating a simulator for finding trustworthy strategies. In an experiment with real users, Kraus et al. (2020b), for example, discovered effects of proactive behaviour on an assistant's perceived trustworthiness depending on the user's technical affinity and domain expertise.
A limitation of presented data collection method is that only the assistant itself used natural language for communication, while the user interacted via actions buttons trigger predefined utterances. Allowing users to interact with the interface using text or even speech input would create another great possibility to capture relevant features (lexical, linguistic, etc.) for predicting the effect of the proactive actions. However, a more complex communication channel would also add noise and increase the possibility of failures, which are independent of the actions and only related to system performance regarding speech recognition and understanding. A more restricted input channel was beneficial for establishing a safe communication between assistant and user. Another drawback was that a perfect system endowed with expert knowledge was used. In a real world scenario, a system that always provides best counseling is unrealistic. However, as the proactive behaviour was randomised and limited to a reasonable frequency, a certain naturalness was added to the agent, as absent system activity could have been perceived as unknowing behaviour. In future work, an error model could be included in the system to simulate not ideal counseling and may be interesting to study, if proactive behaviour remedies a low system performance. A last limitation is the rather sparse data corpus, which needs to be taken into account when applying machine learning algorithms on the data set.

5. Conclusion

In this work, a method for creating a corpus of proactive dialogue was described using a human-to-machine approach. Therefore, an autonomous assistant embedded in a serious game scenario was developed and implemented as a web service. Data from 308 dialogues were collected via crowdsourcing and annotated with several user-dependent, context-dependent, as well as several target variables, whereas the focus was set on the HCT relationship. This forms the first data corpus for the development of proactive dialogue strategies using a rich feature pool. Analysis of the corpus revealed the necessity to consider proactive actions in combination with user characteristics and personality, when developing trustworthy strategies. The corpus is deployed in form of JSON-files and will be available after publication. Additionally, the system's code will be available online and may be extended or altered for individual needs.

Acknowledgments

This work has received funding within the project ”RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots” by the German Federal Ministry of Education and Research.

Abt, C. C. (1970). Serious games. new york: Viking, 1970, 176 pp., $5.95, l.c. 79-83234. *American Behavioral Scientist*, 14(1):129–129.

Balaraman, V. and Magnini, B. (2020). Proactive systems and influenceable users: Simulating proactivity in task-oriented dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, Virually at Brandeis, Waltham, New Jersey, July. SEMDIAL.*

Baraglia, J., Cakmak, M., Nagai, Y., Rao, R., and Asada, M. (2016). Initiative in robot assistance during collaborative task execution. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 67–74. IEEE Press.

Behnke, G., Bercher, P., Kraus, M., Schiller, M., Mickeleit, K., Häge, T., Dorna, M., Dambier, M., Manstetten, D., Minker, W., et al. (2020). New developments for robert–assisting novice users even better in diy projects. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, pages 343–347.

Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Biundo, S. and Wendemuth, A. (2016). Companion-technology for cognitive technical systems. *KI-Künstliche Intelligenz*, 30(1):71–75.

Black, A. W. and Eskenazi, M. (2009). The spoken dialogue challenge. In *Proceedings of the SIGDIAL 2009 Conference*, pages 337–340.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278.*

Chien, S.-Y., Sycara, K., Liu, J.-S., and Kumru, A. (2016). Relation between trust attitudes toward automation, hofstede's cultural dimensions, and big five personality traits. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 841–845. SAGE Publications Sage CA: Los Angeles, CA.

Corti, K. (2006). Games-based learning; a serious business application. *Informe de PixelLearning*, 34(6):1–20.

Crant, J. M. (2000). Proactive behavior in organizations. *Journal of management*, 26(3):435–462.

Evans, A. M. and Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, 42(6):1585–1593.

Gulati, S., Sousa, S., and Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015.

Hoff, K. A. and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434.

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166. ACM.

Isbell, C. L. and Pierce, J. S. (2005). An IP continuum for adaptive interface design. In *Proc. of HCI International.*

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71.

Karrer, K., Glaser, C., Clemens, C., and Bruder, C. (2009). Technikaffinität erfassen–der fragebogen ta-eg. *Der Mensch im Mittelpunkt technischer Systeme*, 8:196–201.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Kraus, M., Fischbach, F., Jansen, P., and Minker, W. (2020a). A comparison of explicit and implicit proactive dialogue strategies for conversational recommendation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 429–435.

Kraus, M., Schiller, M., Behnke, G., Bercher, P., Dorna, M., Dambier, M., Glimm, B., Biundo, S., and Minker, W. (2020b). "was that successful?" on integrating proactive meta-dialogue in a diy-assistant using multimodal cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 585–594, New York, NY, USA. Association for Computing Machinery.

Kraus, M., Wagner, N., and Minker, W. (2020c). Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 107–116, New York, NY, USA. Association for Computing Machinery.

Kraus, M., Wagner, N., and Minker, W. (2021). Modelling and predicting trust for developing proactive dialogue strategies in mixed-initiative interaction. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 131–140.

Kraus, J. M. (2020). *Psychological processes in the formation and calibration of trust in automation.* Ph.D. thesis, Universität Ulm.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909.*

Lowe, R., Pow, N., Serban, I. V., Charlin, L., Liu, C.-W., and Pineau, J. (2017). Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Madsen, M. and Gregor, S. (2000). Measuring human-computer trust. In *11th australasian conference on information systems*, volume 53, pages 6–8. Citeseer.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

McTear, M. F., Callejas, Z., and Griol, D. (2016). *The conversational interface*, volume 6. Springer.

Merritt, S. M. and Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2):194–210.

Merritt, S. M., Heimbaugh, H., LaChapell, J., and Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors*, 55(3):520–534.

Nothdurft, F., Ultes, S., and Minker, W. (2015). Finding appropriate interaction strategies for proactive dialogue systems-an open quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014, Tartu, Estonia*, number 110, pages 73–80. Linköping University Electronic Press.

Paladines, J. and Ramírez, J. (2020). A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*.

Peng, Z., Kwon, Y., Lu, J., Wu, Z., and Ma, X. (2019). Design and evaluation of service robot's proactivity in decision-making support process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 98. ACM.

Rammstedt, B., Kemper, C., Klein, M. C., Beierlein, C., and Kovaleva, A. (2013). Eine kurze skala zur messung der fünf dimensionen der persönlichkeit: Big-five-inventory-10 (bfi-10). *Methoden, Daten, Analysen (mda)*, 7(2):233–249.

Rau, P.-L. P., Li, Y., and Liu, J. (2013). Effects of a social robot's autonomy and group orientation on human decision-making. *Advances in Human-Computer Interaction*, 2013:11.

Ritter, A., Cherry, C., and Dolan, W. B. (2010). Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Sanchez, J., Rogers, W. A., Fisk, A. D., and Rovira, E. (2014). Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science*, 15(2):134–160.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400.

Schmitt, A., Schatz, B., and Minker, W. (2011). Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184.

Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., and Heck, L. (2018). Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Sheridan, T. B. and Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Technical report, Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Statista. (2016). Anzahl der Nutzer virtueller digitaler Assistenten weltweit in den Jahren von 2015 bis 2021, Aug.

Susi, T., Johannesson, M., and Backlund, P. (2007). Serious games: An overview.

Williams, J. D., Henderson, M., Raux, A., Thomson, B., Black, A., and Ramachandran, D. (2014). The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

Yorke-Smith, N., Saadati, S., Myers, K. L., and Morley, D. N. (2012). The design of a proactive personal agent for task management. *International Journal on Artificial Intelligence Tools*, 21(01):1250004.

Young, S., Gašić, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.