Towards Comprehensible Explanations of Phenomena in Home Automation Systems

Matthäus Wander, Viktor Matkovic, Torben Weis, Michael Bischof, Lorenz Schwittmann
University of Duisburg-Essen
Duisburg, Germany

***Abstract*—The current focus in home automation is on making these systems smart and easy to install. Following advances in the area of smart assistants like Alexa and Google Home, we assume that users will not only issue commands to their smart home. They will ask their smart home for explanations why something happened. Hence, we develop and evaluate an algorithm that can explain users why a certain observable phenomenon occured. These questions can originate in the complexity of smart home systems, i.e. the system did something unexpected and the users wonders what caused it. Furthermore, users might ask the system about phenomena caused by their roommates. To evaluate our prototype, we analyze the difference between answers given by humans and those generated by our prototype. Therefore, we conducted an Amazon Mechanical Turk-based Turing Test. In four out of six scenarios our prototype passed the Turing Test. In one of them the computer answer appeared even more human than the real human one.**

I. Introduction

Home automation systems allow for easy control of home appliances. Commercial products based on proprietary technology or on e.g. openHAB [1] are available for purchase, which are simple to install and configure. They support a multitude of sensors and smart devices [2] and are one application of pervasive computing.

With a variety of connected devices and complex control rules, it becomes increasingly hard for humans to comprehend why certain phenomena occur. Consider for example when Debby sits in the kitchen and the coffee machine behind her turns on automatically. Debby would like to understand why this is happening and thus needs a system feature to answer the question: why did the coffee machine turn on? It would be trivial to present the specific rule that has been triggered, for example: because it is between 7 and 8 AM on a weekday and the motion sensor in the staircase activated. However, this is not the answer that Debby is looking for as it is too technical and requires knowledge about the system configuration to comprehend the intention behind this rule. What Debby would like to know, especially if she did not program the system: because Mark is coming downstairs for his morning coffee.

The appropriate level of detail is one dimension of the problem, but another one is the relevancy of events. The phenomenon of a coffee machine turned on can be traced back by following the causal chain of events ultimately to a ringing alarm clock, as this is the first step of Mark's morning routine that involves the coffee machine. But again, the alarm clock is not the information that Debby is looking for.

Existing home automation systems lack this kind of feature. This paper suggests an approach for the generation of human-comprehensible explanations of phenomena in home automation systems. Our approach finds connections between coherent events and extracts relevant information to provide the cause of a phenomenon in a nutshell. This is not only useful for home automation but cyber-physical systems in general, where humans need to grasp quickly the system-triggered behavior without debugging the technical internals. As Norman pointed out [3], the problem with automation is "inappropriate feedback and interaction".

II. System Model

We assume a home automation system with the following key characteristics:

- Phenomena are observed by the system as **events**. They may be the result of **actions** by persons or by the system itself, or natural phenomena like weather events. If a person is involved in an event, the system is capable of recognizing them. This is achieved by sensor fusion [4], which combines low-level data to information such as identifying a person and indoor location tracking.
- A set of '*if this, then that*' **rules** defines the behavior of the system. Rules are logical expressions, which take preceding events as input and trigger an action.
- **Themes** provide thematic context to sensor events, e.g. a rain sensor belongs to 'weather', while a temperature sensor may belong to 'room heating' or 'cooking'. The theme is hard-coded during the set-up of the home automation system.
- The system is capable of tracking ongoing **activities**, which consist of a start event and an explicit end event. For example, watching TV is an activity while entering a room is not.
- **Routines** are series of events, which occur regularly in the same order, e.g. 'Mark wakes up, showers, comes downstairs and takes a cup of coffee'. A routine may consist of multiple activities or events, which happen often consecutively but are otherwise unrelated, e.g. because they affect different rooms or themes. Routines can be either programmed by hand or learned by the system automatically.

The above information are retrievable from a knowledge database. All events are written into a centralized log.

III. Problem Analysis

Prior work has considered information retrieval from e.g. documents and determined that users prefer paragraph-sized chunks of text over an exact phrase in user interfaces [5]. This does not apply to our use case, where the user interacts by speech with the system. Hilton [6] suggested a model that describes how humans give explanations in conversational situations. The problem of giving a comprehensible explanation to a question can be decomposed into two parts: identifying the causes of an event and presenting the relevant information to the person who is asking. The identification of the cause consists of tracing causal connections between events until the most probable cause has been determined. This part depends solely on the knowledge and assessment of the responder, i.e. the system in our case.

Presenting the answer, however, depends on the inquirer because the objective is to give the "most *relevant* answer to the question posed" [6]. The inquirer has a mental model when asking a question, but lacks information to close a gap in his or her knowledge. Information already known to the inquirer should be omitted, as they require attention without providing insight.

Although we cannot replicate the mental model of the individual inquirer accurately with our system, we attempt to approximate a good explanation by considering typically relevant factors. One of the factors is to mention the person that triggered an event, if a person is involved at all. 'Mark entered the room' gives extra information over 'Someone entered the room', which is typically relevant. Mentioning the room or location where the causal event has happened is relevant, if it is another location than the inquirer is at. Time is another factor: the most recent event is usually the most relevant, though older events might be considered with a diminishing relevance as well.

Combining events of different sensors and devices into an overarching theme condenses the relevant information, e.g. 'Mark is cooking' rather than to address which kitchen appliance Mark has just turned on and which device is sensing heat. If a person's action is part of a routine that occurs regularly it is worth mentioning the routine, even though the individual activities within the routine are unrelated, e.g. showering and drinking coffee as part of the morning routine.

IV. Approach

The objective is to give an answer, in which the amount of information is reduced to the relevant pieces. Our approach consists of two parts: 1) construct a causal graph from the system event log, 2) generate an answer by extracting relevant information from the causal graph. Natural language processing is necessary to interact with the user, but out of scope of this paper.

A. Graph Construction

Some of the events stored within the log are causally dependent, which means that a chain of events are set off by a person's or the system's actions. The event log is a flat, time-ordered list. We process the event log and construct a directed, weighted graph, which represents the causality between events that our system asserts. The construction of the graph consists of the following steps: *event partitioning*, *rule matching* and *event linking*.

The following example illustrates the graph construction step by step. An exhaustive presentation of our prototype's algorithm, however, is out of scope of this paper. Consider the following event log:

(1) Debby turns on the TV.
(2) Debby turns off the TV.
(3) Debby enters the kitchen.
(4) Kitchen light turns on.
(5) Debby turns on the oven.
(6) Debby turns off the oven.
(7) Debby leaves the kitchen.
(8) Kitchen light turns off.

From the knowledge database we can derive that events (1) and (2) belong to the activity 'watching TV', events (3)→(4) and (7)→(8) match rules that control the kitchen light, and events (5) and (6) are also an activity and belong to the 'cooking' theme.

The first step is to partition the event log by person and create a separate causal graph for each person. In our example there is only one person involved. Then we look for events that are matching one of the rules of the system. This search happens backwards in the event log, attempting to match actions resulting from a rule and finding preceding input events that match the rule. We thus find rules (3)→(4) and (7)→(8).

Since the log is a time-ordered list and the graph has been grouped by person, the remaining unconnected events can be viewed as a timeline of events that said person participated in. We link the remaining events together and assign a probability of causality as weight to each link. The probability depends on whether the events belong to the same theme, routine or activity, whether they happened in the same room and how far behind in time they are.

B. Answer Generation

Our approach to generate an answer is to traverse the causal graph and extract events and information relevant to a question. Consider for example when Mark asks: '*Why are the lights in the kitchen turned off?*'

We first need to identify the *start event*, which corresponds to the effect mentioned in the question. From there we walk backwards in the causal graph in several phases: *search*, *filter*, *merge themes*, *include involved person* and *include room*. Each phase contributes to the final answer.

We search backwards in time for an event that indicates a finished activity, which serves as cutoff event. The search stops here, unless there are preceding events that are connected with a high probability of causality. This would indicate a potentially relevant routine, thus the search would continue beyond the cutoff event until the probability drops below a threshold. The result is a subgraph on which the remaining

phases rely on. In our example, the start event is event (8) and the cutoff event is the finished activity at event (6).

Afterwards we filter for relevant events. If there is a sequence of multiple events with 100% probability of causality, we select the most recent event and drop the rest of the sequence. The preceding events are not necessary because they always appear together and do not need to be mentioned explicitly. Filtering is not necessary in our example since we have only event (6).

Next, we add the thematic context to events and merge events of the same theme together. Event (6) becomes 'finished cooking' in our example.

Finally, we add the final trigger reason (7), amend the involved person (Debby) and the room (kitchen) to the answer. Thus, the resulting answer to Mark's question about the kitchen light is: *Debby finished cooking* (6) and *Debby left the kitchen* (7).

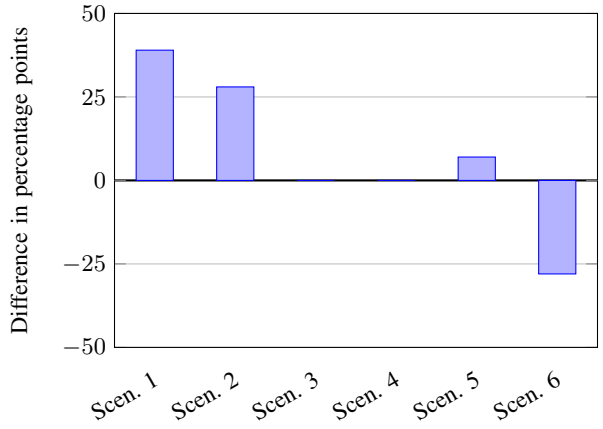

Fig. 1. Difference between human and computer-generated response (ideal: 0%, i.e. not distinguishable).

V. Preliminary Study

For the evaluation of our approach we need to assess whether the answer given by the system is a response useful to humans. In this preliminary study we perform a Turing Test to determine whether humans are able to tell apart the answers of our system from answers that another human would give. The underlying assumption is that humans give useful answers, because they identify the intention of the inquirer intuitively. Although we know that this assumption is not universally true, we aim at gaining insight of what constitutes a good response with this explorative study approach. We chose to recruit test persons from the crowdsourcing platform Amazon Mechanical Turk, because this gives us quick access to a large number of reliable participants. The study consists of two steps:

(1) Our system and test persons give responses in six **example scenarios**.
(2) Test persons rate whether responses from our system and from other persons appear human-like (**Turing Test**).

A. Example scenarios

We created a paper questionnaire with six scenarios, which describe human actions and events observed by a home automation system. Each scenario description is written in natural language and between 29 and 56 words long. Following each scenario, there is a question, in which one of the involved actors asks why something is happening. Test persons were encouraged to give concise answers as if they had to explain it to another person. There was space to write down the answer on one line. We handed out the questionnaire to three test persons, who were unaware of how our algorithm works, and received $3 \cdot 6 = 18$ answers.

In addition to the questionnaire, we modelled the scenarios and let our system prototype generate a response. We consolidate semantically identical answers by ignoring minor grammar differences or the order of words. In some cases the test persons gave identical answers among each other, and in some cases our system generated answers identical to human responses. In total, there are 7 distinct human responses that differ from our system in four out of six scenarios.

B. Turing Test

The objective of the following study is to determine whether our system-generated response can be discriminated from the 7 human responses. We utilize the Amazon Mechanical Turk crowdsourcing platform to recruit test persons and let them decide whether the answer has been given by a human or by an algorithm. The completion of each of the 640 human tasks was compensated with 0.25 US$ each. Each task consists of the scenario description, question and two different answers. For each answer, the test persons can select one of the following options:

- human being
- algorithm
- unclear

We arranged the tasks so that there is always one human and one computer-generated answer (in varying order), but did not inform the test persons about this fact. Thus, the test persons can rate both answers at their own discretion as given by a human or both as given by an algorithm.

If the answer is 'unclear', we assume that the computer-generated answer passes the Turing Test, too, because it is indistinguishable from a human answer. To ensure that test persons do not give arbitrary answers we restricted the participation to workers who had the highest Mechanical Turk approval rating.

Our prototype does not use natural language generation for responses. To avoid that test persons recognize the computer-generated response by language we normalized the answers so that their grammar looks alike and differ only in the *selection* of causal explanations, not in their wording.

C. Results

70 test persons completed 630 out of 640 tasks. We now consider the percentage of answers that were assessed as human-like and compare the percentages between the actual

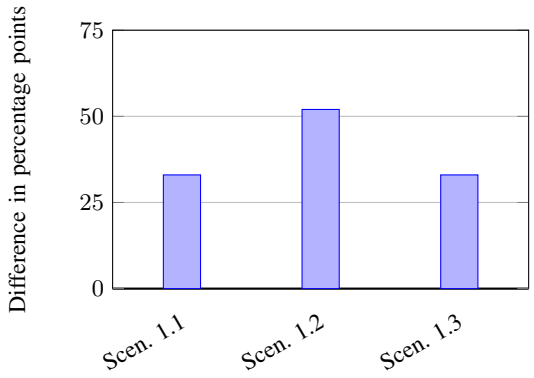

Fig. 2. Detailed comparison for Scenario 1.

human and computer-generated responses. Fig. 1 shows the difference between these two measures: a positive difference indicates that the test persons were on average able to identify what is human and what not, i.e. our system did not pass the Turing Test.

Scenario 3 and 4 were not evaluated as our system generated the same response as humans did, i.e. it passed the Turing Test obviously. The negative difference in Scenario 6 indicates that test persons rated the answers of our system more often as human-like than the actual human answer. When we consider the result of Scenario 5 as minor deviation, our system performed well enough to pass the Turing Test in four out of six scenarios.

We now analyze what caused major differences in the results. Fig. 2 shows the detailed results of **Scenario 1**:

- Question: Why did the coffee machine turn on?
- Human 1.1: Mark wants coffee.
- Human 1.2: Mark programmed it to do so.
- Human 1.3: After showering, Mark wants to drink coffee in the morning.
- Generated: Mark finished showering and is walking down the stairs.

Human responses 1.1 and 1.3 include a conjecture about Mark's intention, which appears more human-like than the factual description in the computer-generated response. Knowing the intention clearly helps to comprehend the system behavior. However, response 1.1 omits that this is a conjecture about Mark's *usual* intention in the morning, which is not necessarily true on every day when the coffee machine turns on. Response 1.3 gives more context and allows to grasp the situation more accurately.

Response 1.2 is interesting as it has been identified clearly as human-like in comparison to the computer-generated response, but does not explain the specific reason that triggered the coffee machine. Instead, the answer reflects that there is a programmed home automation system, which is a generic explanation why a device has turned on. This can be a useful answer if the inquirer is unaware of the home automation system, but with this knowledge in mind it is not.

In **Scenario 2** all three human responses were identical:

- Question: Why did the lights in the study turn on?
- Human: Debby entered the study.
- Generated: Debby arrived at home and is in the study.

The computer-generated response was considered in 50% of all cases as human-like, but the actual human response was considered 78% as human-like, thus the difference. It appears that the information about Debby's arrival is obvious and unnecessary. Thus the algorithm should be adjusted to omit information that can be deduced from other facts.

In **Scenario 6** two test persons and our system gave 'Debby is in the kitchen' as answer, whereas one person responded 'Someone is in the kitchen'. The latter response was assessed as less human-like, which is why our system outperforms the human response. This confirms our thought in Section III that the involved person is typically considered as relevant information.

VI. Conclusion

A good explanation for a phenomenon conveys the relevant information that the asking person lacks and omits information already known or irrelevant. In our preliminary study we conducted a Turing Test to determine what type of answer would be rated as human-like. Our work-in-progress approach has achieved promising results and was not clearly discriminable from human responses in four out of six example scenarios. A key difference to our computer-generated response was that human answers convey the intention of an actor to the inquirer. Thus, future work should attempt to deduce intentions from human actions and existing knowledge. Humans preferred brief responses; information that can be derived from other facts are not worth mentioning.

In order to increase the relevance of an answer, the system should consider the context of the inquirer to adjust the answer subject to the state of knowledge of that person. The mood of that person might be a useful modifier, too. Furthermore, future studies should investigate in-depth not just whether an answer appears human-like, but also whether it is considered helpful.

So far we covered '*why*' questions only. The contrasting '*why not*' questions pose an additional challenge, because they cannot be traced easily to a specific trigger rule.

References

[1] "openhab 2 documentation." [Online]. Available: https://docs.openhab.org
[2] C. Gomez and J. Paradells, "Wireless home automation networks: A survey of architectures and technologies," *IEEE Communications Magazine*, vol. 48, no. 6, 2010.
[3] D. A. Norman, "The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation'," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 327, no. 1241, pp. 585–593, 1990.
[4] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.
[5] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger, "What makes a good answer? the role of context in question answering," in *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*, 2003, pp. 25–32.
[6] D. J. Hilton and B. R. Slugoski, *Conversational Processes in Reasoning and Explanation*. Blackwell Publishers Inc., 2007, pp. 181–206. [Online]. Available: http://dx.doi.org/10.1002/9780470998519.ch9