Designing Natural Language Output for the IoT

Jhim Kiel M. Verame
University of Southampton
Southampton, United Kingdom
j.verame@soton.ac.uk

Enrico Costanza
University of Southampton
Southampton, United Kingdom
ec@ecs.soton.ac.uk

Jacob Kittley-Davies
University of Southampton
Southampton, United Kingdom
jkd3g11@ecs.soton.ac.uk

Kirk Martinez
University of Southampton
Southampton, United Kingdom
km@ecs.soton.ac.uk

Abstract

A large number of devices categorised as "Internet of Things" (IoT) that are in the consumer market are designed to autonomously monitor things of interest to users. These devices often make use of natural language output, more specifically *textual messages*, as a way to notify users. These messages are commonly simple predetermined strings. Some IoT devices however are designed to report on complex applications, which may be difficult for users without technical domain knowledge to understand. In this work, we present an initial evaluation in which we investigated how users' inclination to attend to a monitoring system is affected by different level of information. Based our findings, we discuss future avenues of research which we believe will further our understanding of natural language output's application in the IoT domain.

Author Keywords

natural language output; Internet of Things; user attention

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Abstract

A large number of devices categorised as "Internet of Things" (IoT) that are in the consumer market are designed to au-

tonomously monitor things of interest to users. These devices often make use of natural language output, more specifically *textual messages*, as a way to notify users. These messages are commonly simple predetermined strings. Some IoT devices however are designed to report on complex applications, which may be difficult for users without technical domain knowledge to understand. In this work, we present an initial evaluation in which we investigated how users' inclination to attend to a monitoring system is affected by different levels of information. Based our findings, we discuss future avenues of research which we believe will further our understanding of natural language output's application in the IoT domain.

; H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction

Devices categorised as 'Internet of Things" (IoT) are becoming increasingly available in the marketplace. In particular, a large number of IoT devices are designed to autonomously monitor things of interest to users. Applications range in complexity from simple detect-action applications to the more complex which require human intervention. Skylock [1], a smart bike lock system that alerts users about the potential theft of their bicycle is an example of a detect-action application, it utilise trigger mechanisms to invoke actions which were defined at the configuration phase. An example of a more complex systems where human decision making is required would be the monitoring of livestock [10]. In this situation, IoT devices can inform a veterinarian of a sensed behaviour however in order to make an informed decision, a wealth of knowledge and experience is required.

IoT devices often make use of natural language output, more specifically textual messages, as a way to notify users. In simple applications, predetermined strings are sufficient to convey the necessary information to users. However, how should more complex applications convey sufficient information such that an informed decision can be made by the human recipient without messages being verbose?

We see two key opportunities in exploring how natural language output can be utilised. Firstly how it might be used to convey information from technically complex systems to non-technical users, and secondly how complex messages can be conveyed in a succinct manner to both the technically able and challenged such that the recipient's response is appropriate. In particular in the context of how these messages can attract the appropriate attention of recipients i.e. respond quickly in an emergency or schedule interaction when time allows.

In this paper, we present an initial evaluation in which we investigated how user's attention is affected by different levels of information. In particular, we looked at how users react to messages, shown in different levels of information, containing a report of potential issues produced by a monitoring system. Findings suggest that short non-detailed messages are enough to persuade users to almost always attend to a monitoring system. However, long and detailed messages improves users' efficiency in terms of when to appropriately attend to such systems. Finally, we discuss future avenues of research which we believe will further our understanding of natural language output's application in the IoT domain.

Background

The use the natural language generation in the context of IoT devices is in its infancy and as such there is a limited body of work. As such research from other fields must be considered and built upon.

[1] http://www.skylock.cc/

Recent studies have investigated how familiar metaphors, such as a calendar, can be used to convey complex data to non-technical users (e.g. [2, 11]). In a study by Mennicken et al. [11], they developed and deployed a prototype in two real-world smart homes to examine the effectiveness of a calendar as an interface metaphor to help users make sense of smart home data. Their findings suggest that such a metaphor is helpful in giving an overview of the home's and family's behavioral patterns. In our work, we instead focus on helping users make sense of complex data through synthetic speech.

A number of studies have looked at providing intelligibility to explain to users how certain systems work, in order for users to improve their understanding of the systems' processes (e.g. [1, 5, 7, 8, 9]). For example, in a study by Lim et al. [9], participants were asked to interact with an intelligent system, where each participant received a different explanation of the system's behaviour. Their findings suggest that explaining *why* a system behaved in a certain way helped users the most in their understanding of the system behaviour compared to other types of intelligibility they tested.

Similarly, in a study by Herlocker et al. [5], they evaluated how explanations can improve the acceptance of automated collaborative filtering (ACF) systems. Their findings suggest that participants value having the explanations and that most of the participants also expressed that explanations should be added to ACF systems.

Other studies have also looked at the effects of different levels of system transparency on user's trust (e.g. [3, 4, 6]). For example, in a study by Kizilcec [6], he tested how different levels of system transparency affected users' trust in the context of peer assessment in an online course. Findings suggest that when users' expectations are violated, high transparency (i.e. providing the most detailed explanation) can decrease user's trust in the system. However, when users' expectations are not violated, the different levels of transparency did not affect user's trust.

The studies above focus on how different levels of system transparency can be used to improve user understanding of system behaviour and user's trust in the systems. In contrast, our work focuses on how different levels of system transparency can affect user's attention on issues raised by a monitoring system. In the next section, we present our initial evaluation.

Initial Evaluation

A user study was designed and conducted to understand what level of detail is required in a natural language message system, so that message recipients (users) can make an informed decision about whether to take action or not. Eight messages were generated pertaining to a subject matter which was familiar to all participants, namely server maintenance. We chose this topic as it is sufficiently complex for *some* decisions to take an action (or not) must lie with a human operator. For example, deciding to respond to potential security threats.

Three textual variations were formed for each of the generated messages, with each variation providing more detail than the last. The first variation, “Format A”, expressed a very high level summary of the message being conveyed, the information was limited to one of three predefined categories: User behaviour, System Performance and System Failure. The second message format, “Format B” built upon this with details of the sources which triggered the messages creation, but no specifics in regard of the trigger mechanism. The final message format, “Format C”, provided details of each trigger and the mechanism by which it was invoked. An example of the format can be seen below:

- Format A: “There is a risk of system failure on server *beta*.”
- Format B: “There is a risk of system failure on server *beta* because 1 core metric: temperature, is higher than average.”
- Format C: “There is a risk of system failure on server *beta* because the temperature is 50°C and the average temperature is 9°C.”

The messages were designed such that half of them represented a situation which required users to take action, while the remaining half could be considered “false alarms”. For a false alarm, the example of Format C above would be changed to: “There is a risk of system failure on server *beta* because the temperature is 10°C and the average temperature is 9°C”.

Participants

A total of 8 participants (1 female, 7 male) took part in the study, All of these were members of the University: PhD students and one research assistant, from a variety of disciplines within computer science. All participants have had some experience of networking, server maintenance and troubleshooting.

Method

At the beginning of each experiment, participants were asked to read a participant information document and sign a consent form. They were then given instruction on how to proceed with the task. In more detail, participants were asked to read a scenario text in which they will play a role of a system administrator and would be shown messages about potential technical issues. After showing each message, they were asked what the message meant to them, whether it gave more information than the previous message and what action they would do in response to the message. Participants were placed in a lottery pool to win a £20 in John Lewis voucher. The whole study took around 10 minutes to be completed.

Results and Discussion

Showing format A (the shortest message format) is enough to make people consider taking action such as “investigate what kind of user behaviour it is” (user 1) by for example “checking the user logs” (user 2). However such action may be unnecessary and the comments made by participants reflected that there was insufficient detail to draw a conclusion as “it is not very detailed” (user 3) and that “it’s fairly vague” (user 4). Some participants decided not to act as a result, but the majority expressed a need to investigate further and “try to find out what the risk was” (user 5). It is worth noting that this result may have been caused by the fact that none of the participants are system administrators and instead responded as they would expect system administrators to do so when receiving such a message. In fact, one participant commented that “...if you earn money from this job [as a system administrator], you have to check even if there’s a message saying oh there’s a risk of failure” (user 6).

Message format B did provide enough additional information for the majority of participants to elaborate on their course of action and also increased their likeliness to take an action. However, for most messages, it did not change their initial path even if the situation actually did not need to be investigated. In other words, with both format A and format B participants also reacted to “false alarms”.

In contrast, format C seems to provide the level of detail necessary for most participants to recognize most false alarms, and thus possibly increase their efficiency. All participants reported that Format C (of any message) helped pinpoint the problem with the most precision. As one participant said about Format C: “I think [the message] is detailed. I mean, it makes my life easier. It tells me what I should look at instead of me trying to find the problem. It automatically suggests me what I should check” (user 4).

Future Work

The work outlined here was preliminary and further work is needed to better understand this space. Although our findings suggest that detailed messages result in more decisive decision making, the limited scope of the experiment means that these finding must be only considered advisory and that a more detailed investigation must be conducted to gain further insight.

Having designed this experiment and in reviewing the literature we identified a number of potential avenues for future research. We will refrain from discussing the need to develop a logically and grammatically correct scheme for the generation of messages as it is a field with substantial existing work. However the way in which urgency can be conveyed in the context of inanimate IoT devices could be of great interest. For example a message designed to convey urgency may be misconstrued as low priority. Variation in recipient responses due the the conduit of communication may also be of interest. Does the delivery method impact the actions or recipients? And finally can agent based systems, harnessing two way communication provide natural language systems for IoT devices that not only convey a message, but in a way that invokes the necessary action from the recipient? Also in systems where an autonomous agent can suggest actions, it would be interesting to see how the inclusion of confidence information affects human operator’s actions. In recent work [12] it was shown that the addition of confidence information i.e. estimated probability that an inference that the agent is correct, can improve the adoption and reliance on agents.

Conclusion

In this work, we investigated what level of detail should be included in the natural language messages produced by IoT devices. Our findings suggest different levels of detail have advantages and disadvantages. Displaying short and simple messages are enough to persuade users to attend to the system, which can be more appropriate for safety-critical systems. However, providing detailed messages help users to efficiently take actions, as it helps them distinguish between false and correct alarms. As such, for systems that are less safety-critical, the implementation of detailed messages is more appropriate. Further work is needed to evaluate long term effects and to test the different levels of information in more varied and realistic scenarios.

Acknowledgements

This research was funded, in part, by Innovate UK (ref: 132060). Study approved by U.Southampton FPSE Ethics Committee (ref: 18425).

REFERENCES

1. Bruce G. Buchanan and Edward H. Shortliffe. 1984. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley Series in Artificial Intelligence)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

2. Enrico Costanza, Ben Bedwell, Michael O. Jewell, James Colley, and Tom Rodden. 2016. ’A Bit Like

British Weather, I Suppose': Design and Evaluation of the Temperature Calendar. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4061–4072. DOI: http://dx.doi.org/10.1145/2858036.2858367

3. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455–496.

4. Tove Helldin, Ulrika Ohlander, Göran Falkman, and Maria Riveiro. 2014. Transparency of Automated Combat Classification. In *Engineering Psychology and Cognitive Ergonomics*. Springer, 22–33. DOI: http://dx.doi.org/10.1007/978-3-319-07515-0_3

5. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *CSCW '00: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, 241–250. DOI: http://dx.doi.org/10.1145/358916.358995

6. René F. Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2390–2395. DOI: http://dx.doi.org/10.1145/2858036.2858402

7. Todd Kulesza, Simone Stumpf, Margaret Burnett, Songping Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*. IEEE, 3–10.

8. Brian Y. Lim and Anind K. Dey. 2011. Investigating Intelligibility for Uncertain Context-aware Applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 415–424. DOI: http://dx.doi.org/10.1145/2030112.2030168

9. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128. DOI: http://dx.doi.org/10.1145/1518701.1519023

10. Kevin Mayer, Keith Ellis, and Ken Taylor. 2004. Cattle health monitoring using wireless sensor networks. In *Proceedings of the Communication and Computer Networks Conference (CCN 2004)*. 8–10.

11. Sarah Mennicken, David Kim, and Elaine May Huang. 2016. Integrating the Smart Home into the Digital Calendar. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5958–5969.

12. Jhim Kiel M. Verame, Enrico Costanza, and Sarvapali D. Ramchurn. 2016. The Effect of Displaying System Confidence Information on the Usage of Autonomous Systems for Non-specialist Applications: A Lab Study. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4908–4920. DOI: http://dx.doi.org/10.1145/2858036.2858369