Evaluating Intelligent Knowledge Systems (Article Abstract)

Neil Yorke-Smith[1,2]

[1] Delft University of Technology, The Netherlands
[2] American University of Beirut, Lebanon
n.yorke-smith@tudelft.nl

Abstract. The article published in Knowledge and Information Systems examines the evaluation of a user-adaptive personal assistant agent designed to assist a busy knowledge worker in time management. The article examines the managerial and technical challenges of designing adequate evaluation and the tension of collecting adequate data without a fully functional, deployed system. The PTIME agent was part of the CALO project, a seminal multi-institution effort to develop a personalized cognitive assistant. The project included a significant attempt to rigorously quantify learning capability, which the article discusses for the first time, and ultimately the project led to multiple spin-outs including Siri. Retrospection on negative and positive experiences over the six years of the project underscores best practice in evaluating user-adaptive systems. Through the lessons illustrated from the case study of intelligent knowledge system evaluation, the article highlights how development and infusion of innovative technology must be supported by adequate evaluation of its efficacy.

1 Evaluation of the Personalized Time Management (PTIME) Agent

The case study article by Berry et al [1] reports and critiques the *evaluation* of an intelligent knowledge system that learns preferences over an extended period. The domain of application is personal time management, in particular, providing assistance with arranging meetings and managing an individual's calendar. The *Personalized Time Management* (PTIME) calendaring assistant agent increased in usefulness as its knowledge about the user increases. The enabling technologies involved were preference modelling and machine learning to capture user preferences, natural language understanding to facilitate elicitation of constraints, and constraint-based reasoning to generate candidate schedules [2]. Human-computer interaction (HCI) and interface design played central roles.

The PTIME system was part of a larger, seminal project, *Cognitive Assistant that Learns and Organizes* (CALO), aimed at exploring learning in a personalized cognitive assistant. Thus, the primary assessment of PTIME was in terms of its adaptive capabilities, although such a knowledge-based system must necessarily have a certain level of functionality to assist with tasks in time management,

in order to provide a context for learning. At the commencement of the project, however, the degree of robustness and usability required to support evaluation was not immediately obvious. Evaluation was focused almost exclusively on the technology; experiments were designed to measure performance improvements due to learning within a controlled test environment intended to simulate a period of real-life use—rather than in a genuinely 'in-the-wild' environment. Technologists such as the majority of the authors are trained primarily to conduct such 'in-the-lab' evaluations, but—as argued in the article—many situations require placing the technology into actual use with real users in a business or personal environment, in order to provide a meaningful assessment. In retrospect, the authors suggest that the evaluation methodology of CALO gave too little attention to the usefulness and usability of the technology.

2 Lessons Learned

The six lessons that emerged from the evaluation journey with PTIME are not unfamiliar from other experiences of evaluating (non-adaptive) systems [3]:

1. The contexts of the use of technology, and the competing interests of the stakeholders, must be a primary focus in designing an evaluation strategy.
2. Evaluating one component based on an evaluation of a whole system can be misleading, and vice versa.
3. User-adaptive systems require distinct evaluation strategies.
4. In-the-wild evaluation is necessary when factors affecting user behaviour cannot be replicated in a controlled environment.
5. In-the-wild evaluation implies significant additional development costs.
6. Ease of adoption of the system by users will determine the success or failure of a deployed evaluation strategy.

Summarizing the article, the main lesson from this case study of intelligent knowledge system evaluation is obvious but under-valued: researchers and project managers benefit from familiarity with and adoption of best practice in evaluation methodologies from the start of a technology project.

Acknowledgements This material is based in part upon work supported by the US Defense Advanced Research Projects Agency (DARPA) Contract No. FA8750-07-D-0185/0004. Views are the author(s) and do not necessarily reflect the views of DARPA.

References

1. Berry, P.M., Donneau-Golencer, T., Duong, K., Gervasio, M., Peintner, B., Yorke-Smith, N.: Evaluating intelligent knowledge systems: Experiences with a user-adaptive assistant agent. Knowledge and Information Systems 52, 379–409 (2017)
2. Berry, P.M., Gervasio, M., Peintner, B., Yorke-Smith, N.: PTIME: Personalized assistance for calendaring. ACM Trans. on Intelligent Systems and Technologies 2(4), 40:1–40:22 (2011)
3. Cohen, P., Howe, A.E.: Toward AI research methodology: three case studies in evaluation. IEEE Trans. on Systems, Man, and Cybernetics 19(3), 634–646 (1989)