Dublin Institute of Technology
ARROW@DIT

Conference papers

School of Computing

2009-01-01

Widening the Evaluation Net

Brian Mac Namee
Dublin Institute of Technology, brian.macnamee@dit.ie

Mark Dunne
Dublin Institute of Technology

Follow this and additional works at: http://arrow.dit.ie/scschcomcon

Part of the Artificial Intelligence and Robotics Commons, and the Graphics and Human Computer Interfaces Commons

Recommended Citation
Mac Namee, B. & Dunne, M. (2009) Widening the Evaluation Net, *9th International Conference on Intelligent Virtual Agents (IVA '09)*, Amsterdam, 14-16 Sept.

This Conference Paper is brought to you for free and open access by the School of Computing at ARROW@DIT. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@DIT. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie.
Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License

Widening the Evaluation Net

Brian Mac Namee and Mark Dunne

DIT AI Group, Dublin Institute of Technology, Dublin, Ireland
firstname.lastname@dit.ie

Abstract. Intelligent Virtual Agent (IVA) systems are notoriously difficult to evaluate, particularly due to the subjectivity involved. From the various efforts to develop standard evaluation schemes for IVA systems the scheme proposed by Isbister & Doyle, which evaluates systems across five categories, seems particularly appropriate. To examine how these categories are being used, the evaluations presented in the proceedings of IVA '07 and IVA '08 are summarised and the extent to which the five categories in the Isbister & Doyle scheme are used is highlighted.

1 IVA Evaluations and IVA '08 and IVA '09

As Intelligent Virtual Agent (IVA) research has matured, evaluation has become more important. However, evaluation of IVA systems is notoriously difficult as there are a whole range of issues that must be considered (e.g. *are the behaviours of agents believable?*, *are agents socially capable?*, *does the system run efficiently in real-time?*), and that these issues tend to be quite subjective. However, without good evaluations it is very difficult to compare competing systems and track the development of the field as a whole.

Fortunately, there are a number of proposed standard evaluation schemes for IVA research. One scheme that seems particularly useful was proposed by Isbister & Doyle [1] for evaluating pedagogical conversational agents which evaluates systems under five categories: *Believability*, *Social Interface*, *Application Domains*, *Agency & Computational Issues*, and *Production*.

To examine the state-of-the-art in evaluation in IVA research, the evaluations described in the proceedings of IVA '07 [2] and IVA '08 [3] were summarised. Each full paper published (31 and 45 in IVA '07 and IVA '08 respectively) was examined, and the evaluations described were categorised under the 5 categories in the Isbister & Doyle scheme. Papers for which evaluation is simply inappropriate are placed under the category *N/A*. Finally, those papers that do not describe any evaluations are placed in the category *None*. Figure 1 shows first how many of the papers in each year evaluate under each of the categories in the scheme, and the *N/A* and *None* categories; together with histograms of how many of the categories are covered in the evaluations presented each year.

2 Conclusions & Future Work

The points to notice from the graphs in figure 1 are: there are a large number of papers in which no evaluation is described; it is clear that some of the evaluation

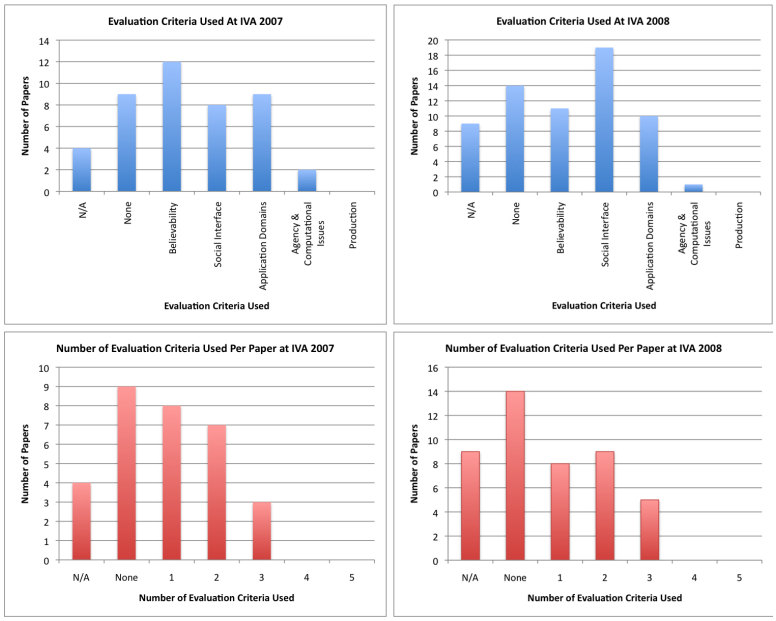

Fig. 1. The number of full papers which evaluated under each of the categories and a histogram of the number of evaluation categories from the Isbister & Doyle scheme used in evaluations reported in full papers at IVA '07 and IVA '08.

categories feature more frequently than others; and in most cases evaluation is performed in only one or two categories. The purpose of this work, so, is to hold a mirror to the evaluations performed within the IVA research community and show that, although there are some example of very fine evaluations reported in the literature, there is still a considerable amount of work to do as the field matures. We would suggest that expanding evaluations to cover better breadth of the Isbister & Doyle scheme would be a good way to move in this direction.

References

1. Isbister, K., Doyle, P.: Design and evaluation of embodied conversational agents: A proposed taxonomy. In: Proceedings of the 1st International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '02) -Conference Workshop: Embodied Conversational Agents -Let's Specify and Evaluate Them. (2002)
2. Pelachaud, C., Martin, J., André, E., Chollet, G., Karpouzis, K., Pelé, D., eds.: 7th International Conference on Intelligent Virtual Agents. Number LNAI4722, Springer (2007)
3. Prendinger, H., Lester, J., Ishizuka, M., eds.: 8th International Conference on Intelligent Virtual Agents. Number LNAI5208, Springer (2008)