NLG Systems Evaluation: a framework to measure impact on and cost for all stakeholders

Cécile Paris [a] **Nathalie Colineau** [a] **Ross Wilkinson** [b]

CSIRO – ICT Centre

[a] Building E6B Macquarie University Campus, North Ryde NSW 2113, Australia

[b] Computer Science & Information Technology Building, ANU Campus, Acton ACT 2601, Australia

{cecile.paris, nathalie.colineau, ross.wilkinson}@csiro.au

1 Enlarging the view of evaluation

The weaknesses of most current evaluation methods is that the conclusions are based not on whether a system performs as expected and on the consequences of its deployment, but on how well it scores against references. In other words, systems are mostly evaluated on some properties (in particular, the "accuracy" of their output), but hardly ever on their ability to fulfil the purpose for which they have been developed and their impact on their (various) users. We argue here that a better way to look at NLG system evaluation would consist in determining the effectiveness of the whole system – not simply its correctness under particular conditions.

Another major drawback of current evaluation practices is to look at only one side of the equation: the benefit. We believe that both the cost and the benefit of the system are important to decide on a system's success.[1] While there is clearly a recognition that there are costs involved, in particular, in obtaining the various resources required (e.g., domain models, task models) – as evidenced by the number of tools developed to help author complex knowledge bases (Power & Scott, 1998; Paris *et al.*, 2005; Androutsopoulos *et al.*, in press) – these costs are typically not measured and not taken into account when evaluating a system. Similarly, the trend towards common evaluation metrics and competitive evaluation tasks does not account for the cost incurred to fine-tune systems for years – a cost also pointed out in (Scott & Moore, 2006). The actual benefit of the improvements may be questionable compared to the cost incurred (e.g., time and effort involved). The benefit-cost tradeoffs (the "bang for buck") are important if we want technology to be adopted and potential users to make an informed choice as to what approach to choose when.

In addition, competitive evaluation tasks often decontextualise systems from their real use by setting artificial tasks. We argue that the context in which a system's effectiveness is evaluated is fundamental – a system exhibiting the 'best performance' might not be the best for a specific task as other task characteristics may be more important.

In this position paper, we consider an NLG system in the context of its stakeholders, their goals and tasks, and the information sources that the system requires. We propose an evaluation framework that allows for all the stakeholders, capturing who benefits from the system and at what cost.

2 A Comparative Framework for Measuring the Effectiveness of NLG Systems

As mentioned in (Paris *et al.*, 2006), and building on work from management and information system, e.g., (Mclean & Delone, 1992; Cornford *et al.*, 1994), we need to enlarge our view of evaluation and identify for each stakeholder role a set of benefits and costs that should be considered. As a first step, we have identified four main stakeholder roles, and, for each, what to evaluate, what questions to ask, as illustrated in Table 1:

– The information *consumer*. The person(s) who will use the generated text.

[1] It might even be useful to look at benefits and costs of a *proposed* system to determine whether it is worth developing and deploying.

	Information Consumer	Information Provider	Information Intermediaries	System Provider
Benefits	Task effectiveness Knowledge gained Satisfaction	Audience reach Audience accuracy Message accuracy	Ease of knowledge creation Ease of context modelling	System usage Reliability Response time Correctness
Costs	Time to complete the task Cognitive load Learning time	Metadata provision Structured information Currency of data	Time to create and integrate the resource Time to capture contextual characteristics	Implementation cost (hardware and software) System maintenance System integration

Table 1. Comparative framework for NLG systems' stakeholders

– The information *provider*. The person(s) (or organisations) with a message to convey. When the generated text is composed of existing text fragments, this person is responsible to provide the content. If the text is generated from first principles, the provider is responsible for the goal(s) and message(s) to be conveyed.

– The information *intermediaries*. They work prior to generation time to create the appropriate set of resources needed by the system (e.g., grammar, lexicon, domain and user models, or potentially text fragments).

– The system providers. They are responsible for the development and maintenance of the technology.

This framework provides us with a context to evaluate different approaches and systems. Given a system (approach) and purpose, the framework forces us to think explicitly about the stakeholders involved, their needs and expectations, how the system meets these and at what cost. This guides us with respect to what experiment(s) to conduct (e.g., test response time or satisfaction of consumers). Ideally, one would want to conduct experiments for each cell in the table. Realistically, we need to identify our priorities for a specific system and carry out the relevant experiments. The results then gives us a way to decide whether the system is worth adopting (developing), given the specified priority(ies) for a given situation (e.g., optimising the benefits to the provider, in particular accuracy of message *vs.* minimising the cost to the intermediary). Note that, the benefits and costs measures might be of a qualitative nature only (e.g., the type of changes required for maintenance and the expertise needed).

When we compare systems within this framework, we do not need the same input and output. What is important is the priority(ies) at stake. In addition, the point is not to average results across the table. Instead, the priorities tell us how to interpret the results. Finally, the framework is not defined around any specific task but can be used to evaluate systems developed for different tasks, given their respective priorities. Note that this approach is whole-of-system oriented.

To conclude, we believe we need to enlarge the view of evaluation, adopting a "consumer-oriented product review" type of evaluation (i.e., whole-of-system), and explicitly thinking of the "bang-for-buck" equation. We have adopted this approach in our own work.

References

Androutsopoulos, I, Oberlander, J., and Karkaletsis, V., in press. Source Authoring for Multilingual Generation of Personalised Object Descriptions. *Natural Language Engineering*, Cambridge University Press.

Cornford, T, Doukidis, G.I. & Forster, D., 1994. Experience with a structure, process and outcome framework for evaluating an information system, *Omega, International Journal of Management Science*, 22 (5), 491-504.

DeLone, W. H. & McLean, E. R., 1992. Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*, 3(1), 60-96.

Paris, C., Colineau, N, Lu S. and Vander Linden, K. 2005. Automatically Generating Effective Online Help. *International Journal on E-Learning*, Vol.4, No.1, 2005. 83-103.

Paris, C., Colineau, N. and Wilkinson, R. 2006. Evaluations of NLG Systems: common corpus and tasks or common dimensions and metrics?. In *Proc. of INLG-06*, held as a workshop on the COLING/ACL Conference, Sydney, Australia, July 15-16. 127-129

Power, R. and Scott, D. 1998. Multilingual authoring using feedback texts. In *Proc. of COLING-ACL 98*, Montreal, Canada.

Scott, D. and Moore, J., 2006. An NLG evaluation competition? Eight Reasons to be Cautious. Technical Report 2006/09, Department of Computing, The Open University.