Evaluations of NLG Systems: common corpus and tasks or common dimensions and metrics?

Cécile Paris, Nathalie Colineau and Ross Wilkinson
CSIRO ICT Centre
Locked Bag 17, North Ryde
NSW 1670, Australia
{Cecile.Paris, Nathalie.Colineau, Ross.Wilkinson}@csiro.au

Abstract

In this position paper, we argue that a common task and corpus are not the only ways to evaluate Natural Language Generation (NLG) systems. It might be, in fact, too narrow a view on evaluation and thus not be the best way to evaluate these systems. The aim of a common task and corpus is to allow for a comparative evaluation of systems, looking at the systems' performances. It is thus a "system-oriented" view of evaluation. We argue here that, if we are to take a system oriented view of evaluation, the community might be better served by enlarging the view of evaluation, defining common dimensions and metrics to evaluate systems and approaches. We also argue that end-user (or usability) evaluations form another important aspect of a system's evaluation and should not be forgotten.

1 Introduction

For this special session, a specific question was asked: what would a shared task and shared corpus be that would enable us to perform comparative evaluations of alternative techniques in natural language generation (NLG)? In this position paper, we question the appropriateness of this specific question and suggest that the community might be better served by (1) looking at a different question: what are the dimensions and metrics that would allow us to compare various techniques and systems and (2) not forgetting but encouraging usability evaluations of specific applications.

The purpose of defining a shared task and a shared corpus is to compare the performance of various systems. It is thus a system-oriented view of evaluation, as opposed to an end-user oriented (or usability) view of evaluation. It is, however, potentially a narrow view of a system-oriented evaluation, as it looks at the performance of an NLG system within a very specific context – thus essentially looking at the performance of a specific application. We argue here that (1), even if we take a system-oriented view of evaluation, the evaluation of NLG systems should not be limited to their performance in a specific context but should take other system's characteristics into account, and that (2) end-user evaluations are crucial.

2 Enlarging the view of system-oriented evaluations

The comparison of NLG systems should not be limited to a particular task in a specific context. Most systems are designed for specific applications in specific domains and tend to be tuned for these applications. Evaluating them in a context of a specific common evaluation task might de-contextualise them and might encourage fine-tuning for this task, which might not be useful in general. Furthermore, the evaluation of a system should not be limited to its performance in a specific context but should address characteristics such as:

- Cost of building (time and effort);
- Ease of extension, maintainability and customisability to handle new requirements (time, effort and expertise required);
- Cost of porting to a new domain or application (time, effort and expertise required);
- Cost of data capture if required (how expensive, expertise required);
- Coverage issues (users, tasks, dimensions of context; and
- Ease of integration with other software.

These dimensions are important if we want the technology to be adopted and if we want poten-

Proceedings of the Fourth International Natural Language Generation Conference, pages 127–129,
Sydney, July 2006. ©2006 Association for Computational Linguistics

tial users of the technology to be able to make an informed choice as to what approach to choose when.

Most NLG systems are built around a specific application. Using them in the context of a different application or domain might be difficult. While one can argue that basic techniques do not differ from one application to another, the cost of the modifications required and the expertise and skills needed may not be worth the trouble. It may be simply cheaper and more convenient to rebuild everything. However, firstly, this might not be an option, and, secondly, it may increase the cost of using an NLG approach to such an extent as to make it unaffordable. In addition, applications evolve over time and often require a quick deployment. It is thus increasingly desirable to be able to change (update) an application, enabling it to respond appropriately to the new situations which it must now handle: this may require the ability to handle new situations (e.g., generate new texts) or the ability to respond differently than originally envisaged to known situations. This is important for at least two reasons:

(1) We are designers not domain experts. Although we usually carry out a domain/corpus/task analysis beforehand to acquire the domain knowledge and understand the users' needs in terms of the text to be generated, it is almost impossible to become a domain expert and know what is the most appropriate in each situation. Thus, the design of a specific application should allow the experts to take on control and ensure the application is configured appropriately. This imposes the additional constraint that an application should be maintainable directly by a requirement specialist, an author, expert or potentially the reader/listener;

(2) Situations are dynamic – what is satisfactory today may be unsatisfactory tomorrow. We must be prepared to take on board new requirements as they come in.

These requirements, of course, come at a cost. With this in mind, then, we believe that there is another side to system-oriented evaluation which we, as designers of NLG systems, need to consider: the ease or cost of developing flexible applications that can be easily configured and maintained to meet changing requirements. As a start towards this goal, we attempted to look more precisely at one of the characteristics mentioned above, the cost of maintaining and extending an application, attempting to understand what we should take into account to evaluate a system on that dimension. We believe asking the following questions might be useful. When there are new requirements:

(1) What changes are needed and do the modifications require the development of new resources, the implementation of additional functionality to the underlying architecture, or both?

(2) Who can do it and what is the expertise required? – NLG systems are now quite complex and require a lot of expertise that may be shared among several individuals (e.g., software engineering, computational linguistics, domain expertise, etc.).

(3) How hard it is? – How much effort and time would be required to modify/update the system to the new requirements?

In asking these questions, we believe it is also useful to decouple a specific system and its underlying architecture, and ask the appropriate questions to both.

3 Usability Evaluations of NLG Systems

When talking about evaluation of NLG systems, we should also remember that usability evaluations are crucial, as they can confirm the usefulness of a system for its purpose and look at the impact of the generated text on its intended audience. There has been an increasing number of such evaluations – e.g., (Reiter *et al.*, 2001, Paris *et al.*, 2001, Colineau *et al.*, 2002, Kushniruk *et al.*, 2002, Elhadad *et al.*, 2005) – and we should continue to encourage them as well as develop and share methodologies (and pitfalls) for performing these evaluations. It is interesting, in fact, to note that communities that have emphasized common task and corpus evaluations, such as the IR community, are now turning their attention to stakeholder-based evaluations such as task-based evaluations. In looking at ways to evaluate NLG systems, we might again enlarge our view beyond reader/listener-oriented usability evaluations, as readers are not the only persons potentially affected by our technology. When doing our evaluations, then, we must also consider other parties. Considering NLG systems as information systems, we might consider the following stakeholders beyond the reader:

- The **creators** of the information: for some applications, this may refer to the person creating the resources or the information required for the NLG system. This might be, for example, the people writing the fragments of text that will be later assembled automatically. Or it might include the person who will author the discourse rules or the templates required. With respect to these people, we might ask questions such as: “Does employing this NLG system/approach save them time?”, “Is it easy for them to update the information?”[1]

- The “**owners**” of the information. We refer here to the organisation choosing to employ an NLG system. Possible questions here might be: “Does the automatically generated text achieve its purpose with respect to the organisation?”, “Can the organisation convey similar messages with the automated system? (e.g., branding issues).

4 Discussion

In this short position paper, we have argued that we need to enlarge our view of evaluation to encompass both usability evaluation (and include users beyond readers/listeners) and system-oriented evaluations. While we recognise that it is crucial to have ways to compare systems and approaches (the main advantage of having a common corpus and task), we suggest that we should look for ways to enable these comparisons without narrowing our view on evaluation and de-contextualising the systems under consideration. We have presented some possible dimensions on which approaches and systems could be evaluated. While we understand how to perform usability evaluations, we believe that an important question is whether it is possible to agree on dimensions for system-oriented evaluations and on “metrics” for these dimensions, to allow us to evaluate the different applications and approaches, and allow potential users of the technology to choose the appropriate one for their needs. In our own work, we exploit an NLG architecture to develop adaptive hypermedia applications (Paris *et al.*, 2004), and some of our goals (Colineau *et al.*, 2006) are to:

- Articulate a comprehensive framework for the evaluation of approaches to building tailored information delivery systems and specific applications built using these approaches.
- Identify how an application or an approach measures along some dimensions (in particular for system-oriented evaluation).

We believe these are equally important for the evaluation of NLG systems.

Acknowledgements

We would like to thank the reviewers of the paper for their useful comments.

References

Colineau, N., Paris, C. & Vander Linden, K. 2002. An Evaluation of Procedural Instructional Text. In the *Proceedings of the International Natural Language Generation Conference (INLG) 2002*, NY.

Colineau, N., Paris, C. & Wilkinson, R. 2006. Towards Measuring the Cost of Changing Adaptive Hypermedia Systems. In *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH2006)*. 259-263, Dublin, Ireland. LNCS 4018.

Elhadad, N. McKeown, K. Kaufman, D. & Jordan, D. 2005. Facilitating physicians' access to information via tailored text summarization. In *AMIA Annual Symposium, 2005*, Washington DC.

Kushniruk, A., Kan, MY, McKeown, K., Klavans, J., Jordan, D., LaFlamme, M. & Patel, V. 2002. Usability evaluation of an experimental text summarization system and three search engines: Implications for the reengineering of health care interfaces. In *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2002).*

Paris, C., Wan, S., Wilkinson, R. & Wu, M. 2001. Generating Personalised Travel Guides? And who wants them? In *Proceedings of the 2001 International Conference on User Modelling (UM’01)*, Sondhofen, Germany.

Paris, C., Wu, M., Vander Linden, K., Post, M. & Lu, S. 2004. Myriad: An Architecture for Contextualised Information Retrieval and Delivery. In *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems (AH2004)*. 205-214, The Netherlands.

Reiter, E., Robertson, R., Lennox A. S. & Osman, L. (2001). Using a randomised controlled clinical trial to evaluate an NLG system. In *Proceedings of ACL'01*, Toulouse, France, 434-441.

[1] We realise that, for some NLG applications, there might be no authors if all the data exploited by the system comes from underlying existing sources, e.g., weather or stock data or existing textual resources.