Task-Based Evaluation of NLG Systems: Control vs Real-World Context

Ehud Reiter
Dept of Computing Science
University of Aberdeen
e.reiter@abdn.ac.uk

Abstract

Currently there is little agreement about, or even discussion of, methodologies for task-based evaluation of NLG systems. I discuss one specific issue in this area, namely the importance of control vs the importance of ecological validity (real-world context), and suggest that perhaps we need to put more emphasis on ecological validity in NLG evaluations.

1 Introduction

Task-based extrinsic evaluation of a Natural Language Generation (NLG) system involves measuring the impact of an NLG system on how well subjects perform a task. It is usually regarded as the 'gold standard' for NLG evaluation, and it is the only type of evaluation which will be seriously considered by many external user communities.

Despite the importance of task-based evaluations, however, there is surprisingly little discussion (or agreement) in the NLG community about how these should be carried out. In recent years there has been a fair amount of discussion about the appropriate use of corpus-based metrics, and there seems (de facto) to be some level of agreement about evaluations based on opinions of human subjects. But there is little discussion and much diversity in task-based evaluation methodology.

In this paper I focus on one one specific methodological issue, which is the relative importance of control and ecological validity (real-world context). An ideal task-based evaluation would be controlled, that is the impact of NLG texts would be compared against the impact of controlled or baseline texts in a manner which minimises confounding factors. It would also be ecologically valid, that is the evaluation would be carried out by representative real-world users in a real-world context while performing real-world tasks. Unfortunately, because of pragmatic constraints including time, money, and ethical approval, it is not always possible to achieve both of these goals. So which is more important?

The methodologies currently used for task-based evaluation in NLG largely derive from the Human-Computer Interaction community, which in turn are largely based on methodologies for experiments in cognitive psychology. Now, psychologists place much more emphasis on control than on ecological validity; they regard control as absolutely essential, but (with some exceptions) they see little wrong with conducting experiments on unrepresentative subjects (undergraduates) in artificial contexts (psychology labs). Indeed many psychologists are now embracing web-based experiments, where they do not even know who the subjects are and what contexts they are working in. For the research goals of psychologists, this probably makes sense. But the research goals of the NLG community are different from the research goals of the psychological community; should we place more emphasis on ecological validity than they do, and less on control?

My own opinions on this matter are changing. Five years ago, I would have echoed the feeling that control is all-important. Now, though, I am beginning to think that in order to achieve both NLG's scientific goals (understanding language and computation) and NLG's technological goals (developing

Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop, pages 28–32,
Edinburgh, Scotland, UK, July 31, 2011.

useful real-world technology), we need to put more emphasis on ecological validity in our evaluations.

2 Evaluation which is both controlled and in real-world context: STOP and DIAG

An ideal evaluation is one which is both controlled and done in a real-world context. An example is the evaluation of the STOP system. which generated tailored smoking-cessation advice based on the user's response to a questionnaire (Lennox et al., 2001; Reiter et al., 2003). The STOP project was a collaboration with medical colleagues, and the STOP evaluation (which was designed by the medics) was carried out as a randomised controlled clinical trial. We recruited 2500 smokers, and sent one-third of them STOP letters, one-third a non-tailored (canned) letter, and one-third a letter which just thanked them for being in out study. After 6 months we asked participants if they had stopped smoking; we tested saliva samples from people who said they had quit in order to verify their smoking status. The result of this evaluation was that the STOP tailored letters were no more effective than the control non-tailored letter. The STOP evaluation cost about UK£75,000, and took about 20 months to design, organise, and carry out.

The STOP evaluation was carried out in a real-world context; the letters were sent to actual smokers, and we measured whether they quit smoking. It was also controlled, since the impact of STOP letters was compared to the impact of non-tailored letters. However there was a lot of 'noise' (in the statistical sense) in the STOP evaluation, because different people (with different personalities, attitudes towards smoking, personal circumstances, etc) received the tailored and non-tailored letters, and this impacted smoking-cessation rates in the the three groups.

Another evaluation which was controlled and was done at least partially in a real-world context was the evaluation of the DIAG-NLP intelligent tutoring system (di Eugenio et al., 2005). In this experiment, 75 students (the appropriate subject group for this tutoring system) were divided into three groups: two groups interacted with two versions of the DIAG-NLP system, and a third interacted with a control version of DIAG which did not include any NLG. Effectiveness was measured by learning gain (change in knowledge, measured by differences in scores in a pre-test and post-test), which is standard in the tutoring system domain. The evaluation showed that students learned more from the second (more advanced) version of the DIAG-NLP system than from the non-NLG version of DIAG.

The DIAG-NLP evaluation was controlled, and it was real-world in the sense that it used representative subjects and measured real-world outcome. However, it appears (the paper is not completely explicit about this) that the evaluation assessed learning about a topic (fixing a home heating system) which was not part of the student's normal curriculum; if this is the case, then the evaluation was not 100% in a real-world context.

3 Evaluation which is controlled but not real-world: BT-45 and Young (1999)

The Babytalk project (Gatt et al., 2009) developed several NLG systems which summarised clinical data from babies in neonatal intensive care (NICU), for different audiences and purposes; one of these systems, BT45 (Portet et al., 2009), summarised 45 minutes of data for doctors and nurses, to support immediate decision-making. Babytalk was a collaborative project with clinical staff and psychologists, and the psychologists designed the BT45 evaluation (van der Meulen et al., 2010).

We picked 24 data sets (scenarios) based on historical data from babies who had been in NICU 5 years previously, and for each data set created three presentations: visualisation, computer-generated text, and human-written text. For each data set, we also asked expert consultants what actions should be taken by medical staff. We then asked 35 medical staff (doctors and nurses of varied expertise levels) to look at the scenarios using a mix of presentations, in a Latin Square design; eg, 1/3 of the subjects saw the visualisation of scenario 1 data, 1/3 saw the computer-generated summary of scenario 1 data, and 1/3 saw the human-written summary of this data. Also each subject saw the same number of scenarios in each condition, this reduced the impact of individual differences between subjects. Subjects were asked to make decisions about appropriate medical actions (or say no action should be taken), and responses were compared to

the 'gold standard' recommendations from the consultants. The result was that decision performance was best with the human-written summaries; there was no significant difference between overall decision performance with the computer-generated summaries and the visualisation (although at the level of individual scenarios, computer texts were more effective in some scenarios, and visualisations was more effective in other scenarios). The BT45 evaluation cost about UK£20,000, and took about 6 months to design, organise, and carry out.

The BT45 evaluation was carefully controlled However, it was not done in a real-world context. Doctors and nurses sat in an experiment room (not in the ward) and looked at data from babies they did not remember (as opposed to babies whom they knew well because they has been looking after them for the past few weeks); they also did not visually observe the babies, which is a very important information source for NICU staff.

Many other task-based evaluations of NLG systems have been controlled but not done in a real-world context, including the very first task-based NLG evaluation I am aware of, by Young (1999). Young developed four algorithms for generating instructional texts, and tested these by asking 26 students to follow the instructions generated by the various algorithms on several scenarios, and measured error rates in carrying out the instructions. The instructions involved carrying out actions on campus (going to labs, playing in soccer matches, etc). The students did not actually carry out these actions, instead they interacted with a 'text-based virtual reality system'. Hence the evaluation was controlled but not carried out in real-world context.

4 Evaluation which is real-world but not controlled: BT-Nurse

The next Babytalk system (after BT45) was BT-NURSE; it generated summaries of 12-hours of clinical data, to support nursing shift handover (Hunter et al., 2011). We initially expected to evaluate BT-NURSE using a similar methodology to the BT45 evaluation. However the medical people involved in BabyTalk complained that it was unrealistic to evaluate the system in an artificial controlled context, where clinical staff were looking at data out of context. So instead we evaluated BT-NURSE by installing the system in the NICU, so that nurses used it to get information about babies they were actually caring for. The primary outcome measure was subjective ratings by nurses as to the helpfulness of BT-NURSE texts; and indeed most nurses thought the texts were helpful.

The BT-NURSE evaluation was significantly more expensive than the BT45 evaluation, because we hired a full-time software engineer for a year to ensure that the software was sufficiently well engineered so that it could be deployed and used in the hospital; we were also required by the medical ethics committee to have a research nurse on-site who checked texts for errors before they were shown to the duty nurses, and removed them from the experiment if they were factually incorrect and could damage patient care (in fact this never happened, the research nurse did not regard any of the BT-NURSE texts as potentially harmful from this perspective). All in all cost was probably about UK£50,000, and the entire process (including the software engineering) took about 18 months.

The BT-NURSE evaluation was not controlled; we did not compare the computer generated texts to anything else, and indeed did not directly measure any task outcome variable, instead we solicited opinions as to utility. It was however ecologically valid, since it was carried out by asking nurses (real-world users) to use BT-NURSE for care planning (real-world task) in a real-world context (on-ward, involving babies the nurses were familiar with and could visually observe).

5 Discussion

Ideally a task-based evaluation should be both controlled and ecologically valid (done in a real-world context). But if it is not possible to achieve both of these objectives, which is most important? Obviously in many cases the desires of collaborators need to be considered; for example psychologists generally place much more emphasis on control than on ecological validity, whereas many commercial organisation take the opposite perspective. But which is more important from an NLG perspective?

From a pragmatic perspective, two important arguments for focusing on control are cost and publi-

cations. The figures given above suggest that doing an evaluation in a real-world context makes it substantially more expensive. Of course this is based on very limited data, but I believe this is correct, deploying a system in a real-world context requires addressing engineering and ethical issues which are expensive and time-consuming to resolve. From a publications perspective; most NLG reviewers are much more concerned about control than about ecological validity. Especially in high-prestige venues, reviewers are likely to complain about uncontrolled evaluations, while making little (if any) mention of concerns about lack of ecological validity.

For what its worth, my own view on this issue has changed. If asked five years ago, I would have said that control was more important, but now I am veering more towards ecological validity. The technological goal of NLG is to develop technology which is used in real-world applications, and from this perspective if we do not evaluate in real-world contexts, we risk being side-tracked into technology which looks good in a controlled environment but is useless in the real world. Similarly, if our goal is to develop a better scientific understanding of computation and language, I think we have to look at how language is used in real-world contexts, which (at least in my mind) is quite different from how language is used in artificial contexts.

Plaisant (2004) made some related points in her discussion of evaluation of information visualisation. She pointed out that controlled evaluations of visualisation systems in artificial contexts might be less informative than uncontrolled evaluations in real-world contexts. She also pointed out that controlled evaluations could not evaluate some of the most important benefits of visualisation systems. For example, sometimes the primary objective of visualisation systems is to support scientific discovery, that is to make it easier for scientists who are analysing data to come up with new insights and hypotheses. However, testing effectiveness at supporting scientific discovery in a controlled fashion is almost impossible. Perhaps in theory one could compare the 'productivity' of two groups of scientists, one with and one without visualisation tools, but the comparison would have to involve a large number of scientists over a period of months or even years, with scientists in one group not allowed to communicate with scientists in the other group. It is difficult to imagine that such an experiment could in fact be carried out (or that it would be approved by a research ethics committee). Plaisant argues that focusing on controlled experiments means focusing on things that are easily measurable in such experiments, which may lead researchers to ignore the outcomes that we really care about.

Another important point is that the goal of evaluation is not just to assess if something works, but also to come up with insights as to how to improve an algorithm, module, or system. In NLG evaluations such insights are often based on free-text comments made by subjects, and in my experience better and more insightful comments are obtained from evaluations in real-world contexts.

An important potential caveat is that all of the examples cited above were system evaluations, which attempted to assess how useful a system was from an applied perspective. If the goal of an evaluation is to test a scientific theory or model, should we always (as psychologists do) favour control over ecological validity? My own belief is that the psychologists are missing important insights and findings by ignoring ecological validity, and the most effective way for the NLG community to 'add value' to the enterprise of understanding language is not to imitate the psychologists, but rather to use a different experimental paradigm, which focuses much more on ecological validity. But others will no doubt disagree.

6 Conclusion

It is difficult to choose between control and ecological validity, because clearly both greatly contribute to the usefulness of an evaluation. But this tradeoff must be made in many cases, and it would be preferable for it to be explicitly discussed. And of course there are many other desirable factors which may need to be involved in a tradeoff; for example, how important is it that subjects be representative of the user community, instead of whoever is easiest to recruit (eg, undergraduates). My hope is that the NLG community can explicitly discuss such issues, and come up with recommended evaluation methodologies for task-based studies, which are based the scientific and technological objectives of our community.

References

Barbara di Eugenio, Davide Fossati, Dan Yu, Susan Haller, and Michael Glass. 2005. Aggregation improves learning: experiments in natural language generation for intelligent tutoring systems. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics.

Albert Gatt, Francois Portet, Ehud Reiter, Jum Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications*, 22(3):153–186.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, Cindy Sykes, and Dave Westwater. 2011. BT-Nurse: Computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal of the Americal Medical Informatics Association*. In press.

Scott Lennox, Liesl Osman, Ehud Reiter, Roma Robertson, James Friend, Ian McCann, Diane Skatun, and Peter Donnan. 2001. The cost-effectiveness of computer-tailored and non-tailored smoking cessation letters in general practice: A randomised controlled study. *British Medical Journal*, 322:1396–1400.

Catherine Plaisant. 2004. The challenge of information visualization evaluation. In *Proceedings of Advanced Visual Interfaces (AVI) 2004*.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Ehud Reiter, Roma Robertson, and Liesl Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

Marianne van der Meulen, Robert Logie, Yvonne Freer, Cindy Sykes, Neil McIntosh, and Jim Hunter. 2010. When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology*, 24(1):77–89.

Michael Young. 1999. Using Grice's maxim of quantity to select the content of plan descriptions. *Artificial Intelligence*, 115:215–256.