NLG Shared Tasks: Lets try it and see what happens

Ehud Reiter
Department of Computer Science
University of Aberdeen
Aberdeen AB24 3UE, UK
ereiter@csd.abdn.ac.uk

1 Pros and Cons of Shared Tasks

I must admit that I have mixed feelings about shared task evaluations. Shared task evaluations of course have many advantages, including allowing different algorithms and approaches to be compared, producing data sets and evaluation frameworks which lower the "barriers to entry" to a field, and more generally getting researchers to interact more, and realise how their assumptions about inputs, outputs, knowledge sources, and processing constraints differ from those made by other researchers.

Shared task evaluations could also help us understand evaluation better. I would like to get a better idea of how different evaluation techniques (such as statistical evaluation, human preference judgements, and human task performance) correlate with each other. In order to carry out such studies, it would be very useful to have a number of systems with similar input/output functionality and knowledge sources; a shared-task evaluation could provide these systems (Reiter and Belz, 2006).

On the other hand, there are also dangers to shared tasks. In particular, focusing on a shared task can cause a community to narrow the scope of what it investigates. For example, colleagues of mine in the Information Retrieval community have suggested to me that the academic IR community's focus on the TREC shared evaluation in the mid and late 1990s limited its contribution to web search when this emerged as the "killer app" of IR. This is because the 1990s academic IR community had little interest in web-search algorithms (such as Google's page rank) which could not be used in TREC shared tasks.

In other words, TREC encouraged the IR community to focus on one specific type of IR problem, and probably helped it make progress in this area. But this was at the cost of ignoring other types of IR problems, which turned out to be more important.

My personal opinion is that we should try to organise some shared task evaluations in NLG, but do this (at least in the first instance) as one-off exercises. I think a yearly "NLGUC" event would be a mistake; but I think one-off shared evaluations could be worthwhile and should be tried.

2 Issue: Topic

From a practical perspective, I suspect that the main challenges in running an NLG shared evaluation are going to be (1) choosing a topic that attracts enough participants to make the exercise meaningful, and (2) deciding how to evaluate systems. Looking at the topic issue first, the NLG community is quite small (recent International NLG conferences have attracted on the order of 50 people), and the NLG problem space is enormous. Since a shared task evaluation must focus on specific NLG problem(s), it is not easy to find a topic which would attract a reasonable number of participants (at least 6, say).

One possible topic that could attract this number of people is generating referring expressions. This has attracted a lot of attention in recent years; for example in INLG 2006 there were papers on this topic from groups in Australia, Brazil, Germany, Japan, UK, and USA. There are also some corpora available which could be used for a reference generation shared task, such as Coconut (Jordan and Walker, 2005) and the Tuna corpus (van Deemter et

al., 2006).

Another possibility, which focuses on an application instead of on an NLG task, is generating weather forecasts. This has been one of the most popular NLG applications over the past 20 years; Bateman and Zock's list of NLG systems[1] (which is not complete) lists 13 systems in this area. And there are corpora available, such as the SumTime corpus (Sripada et al., 2005).

A third possibility is medical, in particular patient information. Medical applications of NLG are popular according to Bateman and Zock's list, and there are many people outwith the NLG community who are interested in generating personalised health information; indeed there are workshops on this topic. However, I suspect it would be harder to organise a shared task evaluation in this area because data resources would need to be created (I'm not aware of any existing corpora in this area).

3 Issue: Evaluation

Another challenge in organising a shared task evaluation is deciding how to evaluate the systems. I believe that most shared task evaluations in Language Technology use corpus-based evaluation, but this can be controversial, not least because corpus-based evaluation metrics seem to be biased towards systems built using corpus-based techniques (Belz and Reiter, 2006). In NLG in particular, it is clear that writers do not always produce optimal texts from the perspective of readers (Oberlander, 1998; Reiter and Sripada, 2002); this is another argument against using metrics which compare machine-generated texts to human written texts.

But reader-based evaluations have problems as well. The easiest kind to carry out is rating exercises, where human subjects are asked to rate the quality of generated texts. However, we know that in many cases such ratings are not good predictors of how useful texts actually are in helping real users carry out real tasks (Law et al., 2005). Task-based evaluations are more robust in this sense, but they are expensive and time-consuming, and we have no guarantees that texts that are useful in supporting one task will also be useful in supporting other tasks.

[1] http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm

Given this uncertainty, I think any shared task evaluation should use a number of different evaluation techniques. Indeed, as mentioned above, I think one of the goals of a shared task evaluation should be to get empirical data on how well different evaluation metrics correlate with each other, so that discussions about evaluation techniques can be informed by real data.

The other advantage of multiple evaluation techniques is that it makes it harder to say who "won" a shared task evaluation. This is good, because I think the NLG community will be more willing to participate in shared task evaluations if they are primarily seen as scientific ventures instead of as contests.

References

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *Proceedings of EACL-2006*, pages 313–320.

Pamela Jordan and Marilyn Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Anna Law, Yvonne Freer, Jim Hunter, Robert Logie, Neil McIntosh, and John Quinn. 2005. Generating textual summaries of graphical time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.

Jon Oberlander. 1998. Do the right thing ... but expect the unexpected. *Computational Linguistics*, 24:501–507.

Ehud Reiter and Anja Belz. 2006. Geneval: A proposal for shared-task evaluation in nlg. In *Proceedings of INLG 2006*.

Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of the Second International Conference on Natural Language Generation*, pages 97–104.

Somayajulu Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2005. SUMTIME-METEO: Parallel corpus of naturally occurring forecast texts and weather data (revised 2005 edition). Technical Report AUCS/TR0201, Computing Science Dept, Univ of Aberdeen, Aberdeen AB24 3UE, UK.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of INLG 2006*.