Introduction to the INLG'06 Special Session on Sharing Data and Comparative Evaluation

The idea for this special session had its origins in discussions with many different members of the NLG community at the 2005 Workshop on Using Corpora for Natural Language Generation (UCNLG'05, held in conjunction with the Corpus Linguistics 2005 conference at the University of Birmingham in July 2005), and subsequently at the 10th European Natural Language Generation Workshop (ENLG'05, held at the University of Aberdeen in August 2005). At the latter event, the excitement about introducing shared tasks was infectious: the topic hijacked several of the organised discussion groups, it was the focus of conversation at many tables during lunch-breaks, and even the end of the conference didn't put a stop to it, with discussions carrying right on until the taxis to the airport arrived.

There was some common ground: nobody said it wouldn't be a good idea to be able to directly compare different NLG techniques, most people even seemed to agree that sharable data and tasks were the way to go. But opinion was sharply divided about how it was to be achieved. There were—with only a small degree of caricature—two main camps: the bulls argued for a suck-it-and-see approach, for throwing a task at the research community and then sitting back to see what would happen; the bears warned that using one data set was not a good idea until there was community buy-in to a particular data set and a particular task specification over that data set.

Some of the bears were worried that NLG would inevitably emulate MT and end up with a single task, fixed inputs and gold-standard outputs, using a single automatic metric of similarity to assess the quality of generated texts against the gold standard, and moreover, require millions of dollars in direct funding. It would be impossible to decide what the inputs to the task should look like, because after all, as Yorick Wilks had pointed out years before, determining the inputs to NLG was like counting back from infinity to 1 (in contrast to NLU, which, being more akin to counting from 1 to infinity, seemed at least a little more manageable). The community would either become hopelessly mired in the task of trying to agree on an input type and task, or else agree one by dictat and alienate the majority of researchers. Finally, the field would become obsessed with the single task and the scores produced by the single metric, and all true scientific enquiry would be stifled.

The bulls envisaged an entirely different future, where many different tasks and benchmark datasets co-existed peacefully, where some tasks did have associated inputs and outputs, but others had more abstract system specifications. NLG was not inherently different from NLU at all, in fact the output representations used in the latter were just as much there by gentle(wo)man's agreement as any common inputs to NLG would be. The strong NLG traditions of user-oriented and task-based evaluations using human evaluators would be part of the evaluation paradigm in shared-task evaluations, while parallel research might look at—but not impose—bespoke automatic methods for NLG. Money would be needed for data resource creation, but not necessarily for anything else; evidence that this was possible could be found in successful and vibrant shared-task initiatives run on a shoe-string, such as CoNLL and SENSEVAL. The community would create its own forum for reviewing, updating and adding tasks and evaluation methods. NLG would be invigorated, great scientific progress would result, commercial deployment of NLG technology and regular papers in *Computational Linguistics* and ACL proceedings would surely follow.

One thing was clear: opinions abounded, most of them strong ones. Shared-task evaluation had been firmly put on the NLG agenda. So, we thought, what better than to create a larger,

Proceedings of the Fourth International Natural Language Generation Conference, pages 125–126,
Sydney, July 2006.

more enduring forum for continuing discussions, in the shape of an INLG special session? We are pleased to say that the response from the NLG community has been very positive, and that the papers in this section of the proceedings and the oral presentations at the special session itself represent both the bullish and the bearish camps. Belz and Kilgarriff present a generally bullish, but occasionally bearish, history of shared-task initiatives in NLP, and the lessons that NLG might learn from it, while Reiter and Belz present a proposal for a series of shared tasks in data-to-text NLG. Van Deemter et al. look at the generation of referring expressions and propose a method for eliciting reference texts for evaluation of GRE algorithms. Paris argues for NLG system evaluation practices similar to the ISO standards for software evaluation, including criteria such as flexibility, portability and maintainability.

Among the oral presentations, Scott and Moore exemplify the bearish position but do argue in favour of a standardised architecture and interface specifications to eventually enable cross-system comparison. Horacek considers the input problem and advocates the gradual and collective development of a generic 'generation specification' formalism. Varges recommends that NLG deliberately take a different route from NLU and encourage a diversity of tasks and representations.

Kathy McKeown's invited talk is perfectly poised between the two camps: from her experience with DUC, TREC and GALE, she concludes that every evaluation programme must expect to have to weather a stormy period of initial disagreement and even hostility, before eventually reaching calmer waters where growing agreement and acceptance enable the true benefits of the programme to take effect.

Consensus-spotters will be able to identify several areas of interest: certainly nobody wants to follow the example of MT and parsing, and become beholden to a single metric and automated gold-standard evaluation; some degree of standardisation in sub-tasks and representations is desirable, but should evolve over time; and perhaps most unanimously, the diversity of current NLG research with its many different tasks and interests must be preserved.

Even a small amount of common ground can be enough for debate to flourish and consensus to grow. We hope that the snapshot of opinion presented at this special session will be the beginning of a long history of comparative evaluation in NLG.

Anja Belz and Robert Dale (Organisers; one bull and one bear)