An NLG evaluation competition? Eight Reasons to be Cautious

Donia Scott
Centre for Research in Computing
The Open University, U.K.
FirstInitial.Lastname@open.ac.uk

Johanna Moore
Human Communication Research Centre
The University of Edinburgh, U.K.
FirstInitial.Lastname@ed.ac.uk

Abstract

It is our view that comparative evaluation of the type used in MUC, TREC, DUC, Senseval, Communicator, may not be sensible for NLG and could be a misguided effort that would damage rather than help the field.

Most would agree that NLG has to date failed to make as significant impact on the field of NLP and on the world–as measured by the number publications, existing commercial applications, and the amount of funding it has received. While it may be useful to look at other subfields of NLP (e.g., message understanding, machine translation, summarization, word sense disambiguation) and speculate why this should be the case, we urge caution in proceeding under the assumption that a good path to progress in NLG would be to jump on the evaluation competition bandwagon.

All that glitters is not gold: For evaluation competitions to have much meaning, there has to be a gold standard to aspire to. With a clearly defined input and a fully-specified output, one may be able to establish a reasonable criterion for success that can be applied to all competitors. In the case of NLG, this is extremely hard to achieve–some may say impossible–without distorting the task to a degree that renders it otiose.

What's good for the goose is not necessarily good for the gander: NLG systems have been, and continue to be built to serve a wide range of functions. It makes little sense to compare the output of systems that are designed to fulfull different functions, especially since the most important criterion for any system is its "fitness for purpose". NLG, unlike MT and parsing, is not a single, well-defined task but many, co-dependent tasks.

Don't count on metrics: Both the summarization and the MT communities, who have for several years been working towards shared metrics, are now questioning the usefulness of the metrics. For the past 3–4 years, to claim that one has made progress in MT, one simply needed to report an increase in BLEU score. Yet in the past year, there have been several papers published decrying the usefulness of BLEU (e.g., Callison-Burch *et al.* (2006)), and showing that it does not correlate well with human judgements when it comes to identifying high quality texts (despite prior reports to the contrary). Indeed, the recent word on the street is that BLEU should only be used as one of many metrics to tell if one is improving their own system, *not* as a metric to compare systems (Kevin Knight, invited talk, EACL 2006). Simply put: so-called 'quality metrics' often don't give you what you want, or what you think they give.

What's the input? The difference between NLU and NLG has been very aptly characterised as the difference between counting from one to infinity or from infinity to one (Yorick Wilks, invited talk, INLG 1990). A huge problem in NLG is that, quite simply, different applications have different input. But, even if we were to agree on a shared

task (and this is a huge problem in itself) such as producing reports of stock market activity, some would advocate starting with the raw data coming off the ticker tape, while others would say that the data analysis program needed to identify significant events in the data stream has nothing to do with the generation process. But surely the quality of data analysis will affect the quality of the text that is generated.

What to standardize/evaluate? So what can we hope to provide evaluation metrics for? Some would argue that realization is clearly an area for which we can provide standard metrics because surely we can all agree on what the input and output specification should be. But even here, there will be heated debate not only over what formalism to use, but what information must be specified in the input. For example, should the input to the realizer be required to include information structure? Should the output include markup for pitch accents and boundary tones (which is needed for high-quality speech synthesis)? If information structure is essential to your theory of how many generation choices are made, you will argue vehemently for it. But if it does not fit your theory or you don't have a content and sentence planner capable of producing the semantically rich input representation required, you will argue just as vehemently against it.

The plug-and-play 'delusion': One of the main selling points of the DARPA Communicator program was the idea of plug-and-play. It was intended to give researchers a full end-to-end dialogue system, in which they could test competing hypotheses about one component of a system (e.g., the parser, the dialogue manager, the response generator) without building all the other components. Great idea; horrific execution. Communicator specified a low-level agent communication architecture (Galaxy Communicator), *not* the interfaces between components of a dialogue system. The result was that the plug-and-play dream never came to fruition. And despite a large scale NIST evaluation of nine systems all performing the same task, many would claim that the dialogue community has learned virtually nothing about how to build better dialogue systems from this time-consuming and expensive exercise.

Who will pay the piper? The reason that ATIS, Communicator, BLEU, ROUGE, DUC, TREC, etc., made it past the coffee room is literally *millions* of U.S. dollars of research funding. If NLG hopes to get any momentum behind any evaluation initiative, there has to be a funder there to pay the bills. Who will do this, and why should they? Put another way: what's the 'killer app' for NLG in the Homeland Security domain?

Stifling science: To get this off the ground we have to agree the input to realization. And you can push this argument all the way up the NLG pipeline. And whatever we agree on will limit the theories we can test. So what is really needed is a theory neutral way of representing the subtask(s) of the generation process to be evaluated. If we cannot do this, we will stifle new and truly creative ideas that apply new advances in linguistics to the generation process.

We believe that a good starting point in being able to compare, evaluate and maybe even reuse NLG technologies could be for the community to engage with something like the RAGS initiative, which provides a language for describing the interfaces between NLG components (Mellish et al., 2006). We also think that the NLG community would benefit from becoming better versed in the experimental methods for conducting human evaluation studies. Until then, there is a real risk that too many people will engage in wasted efforts on invalid or irrelevant evaluation studies, and some good but unsexy evaluation studies will continue to be misunderstood.

References

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

C. Mellish, D. Scott, L. Cahill, D. Paiva, R. Evans, and M. Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12:1–34.