

An NLG evaluation competition? Eight Reasons to be Cautious

Donia Scott

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

Johanna Moore

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

Most would agree that NLG has to date failed to make much of an impact on the field of NLP and on the world — as measured by the number of articles in *Computational Linguistics*, papers in ACL conferences, and existing commercial applications, and by the amount of funding it has received. While it may be useful to look with envy at other subfields of NLP (e.g., message understanding, machine translation, summarization, word sense disambiguation) and speculate why this should be the case, we would urge for caution in proceeding under the assumption that a good path to progress in NLG would be through jumping 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; for example:

- generating responses in dialogue (Stent et al., 2002; Walker et al., 2002; Moore et al., 2004);
- drafting technical documents (Paris et al., 1995; Power and Cavallotto, 1996; Hartley et al., 2001);
- drafting weather reports (Goldberg et al., 1994; Coch, 1998; Vavargard,

2000);

- producing captions for diagrams (Mittal et al., 1998; Fasciano and Lapalme, 2000);
- editing knowledge-bases (Power and Scott, 1998; Scott et al., 1998);
- querying databases (Waltz, 1978; *et al.*, 1978; Hallett et al., 2005);
- writing letters (Springer et al., 1991; Coch et al., 1995; Reiter et al., 2003);

to name but a few. It makes little sense to compare the output of systems that are designed to fulfill different functions, especially since the most important criterion for any system is its “fitness for purpose”. The bold fact is that 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. If we want to produce reports of stock market activity, we might agree that we should use the raw data coming off the ticker tape, or we may agree that we have some data analysis package that produces significant events to be reported in some agreed format. But then we have to do this for every domain. The fact of the matter is that content structuring and organization is always going to be domain and task specific. So, we'll need to come up with an agreed task for NLG — and that in itself is hugely problematic.

What to standardize/evaluate? So what can we hope to be providing evaluation metrics for? Some would argue that realization is clearly an area for which we can provide standard metrics because surely we know the input and output that we expect? OK, if you agree with that, then consider what you would say if the input specification we agreed upon were Hybrid Logic Dependency Semantics (which includes information structure) or Minimal Recursion Semantics, and the output had to include markup for pitch accent and boundary tones (which are needed for high-quality speech synthesis)? You might say, we're only interested in text. But first, who's "we"? Many of us are interested in producing spoken language output. And second, many will argue that information structure is necessary for a principled approach to many kinds of coordination, even in text. And, of course, others won't want to deal with it because it does not fit their theory or they don't have the content and sentence planners that are capable of producing the semantically rich input representation required.

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 systems, 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 will limit the theories we can test. So what is really needed is a theory neutral way of representing the subtask of the generation process that is 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). However, while RAGS has been a step in the right direction, it still falls short of being a final solution, in that its interfaces do not include information structure, and therefore limit our ability to produce theory-neutral descriptions. 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*.
- J. Coch, R. David, and J. Magnoler. 1995. Quality test for a mail generation system. In *Proceedings of Linguistic Engineering'95*, pages 435–443, Montpellier, France.
- J. Coch. 1998. Interactive generation and knowledge administration in multimeteo. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 300–303, Niagara-on-the-lake, Ontario, Canada. Demonstration.
- E. F. Codd *et al.* 1978. Rendezvous version 1: An experimental English-language query formulation system for casual users of relational databases. Technical Report RJ2144(29407), IBM Research Laboratory, San Jose, Calif.
- M. Fasciano and G. Lapalme. 2000. Intentions in the coordinated generation of graphics and text from tabular data. *Knowledge and Information Systems*, 2(3).
- E. Goldberg, N. Driedger, and R. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.
- C. Hallett, D.Scott, and R. Power. 2005. Intuitive querying of e-health data repositories. In *Proceedings of the UK E-Science All-Hands Meeting*, Nottingham, U.K.
- A. Hartley, D. Scott, J. Bateman, and D. Dochev. 2001. AGILE — a system for multilingual generation of technical instructions. In *Proceedings of the Machine Translation Summit VIII*, pages 145–150, Santiago de Compostela, Spain.
- V. O. Mittal, J.D. Moore, G. Carenini, and S. Roth. 1998. Describing complex charts in natural language: a caption generation system. *Computational Linguistics*, 24(3):431–468, September.
- J. Moore, M.E. Foster, O. Lemon, and M. White. 2004. Generating tailored, comparative descriptions in spoken dialogue.
- C. Paris, K. Vander Linden, M. Fischer, A. Hartley, L. Pemberton, R. Power, and D. Scott. 1995. A support tool for writing multilingual instructions. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1398–1404, Montréal, Canada, August.
- R. Power and N. Cavallotto. 1996. GIST: Multilingual generation of administrative forms. In *INLG'96*, pages 17–19, Herstmonceux Castle, Sussex. Demonstration.
- R. Power and D. Scott. 1998. Multilingual authoring using feedback texts. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 1053–1059, Montreal, Canada.
- E. Reiter, R. Robertson, and L. Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.
- D.R. Scott, R. Power, and R. Evans. 1998. Generation as a solution to its own problem. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 256–265, Niagara-on-the-Lake, Ontario.
- S. Springer, P. Buta, and T. Wolf. 1991. Automatic letter composition for customer service. In Reid Smith and Carlisle Scott, editors, *Innovative applications of artificial intelligence 3*. AAAI Press. (Proceedings of CAIA-1991).
- A. Stent, M. Walker, S. Whittaker, and P. Maloor. 2002. User-tailored generation for spoken dialogue: An experiment. In *Proceedings of ICSLP*.
- T. Vavargard. 2000. Autotext. In *Preprints 2nd Conference on Artificial Intelligence*, pages 69–71, Long Beach, CA. American Meteorological Society.
- M. Walker, O. Rambow, and M. Rogati. 2002. Training a sentence planner for spoken dialog using boosting. *Computer speech and language*, 16:209–433.
- D. L. Waltz. 1978. An English language question answering system for a large relational database. *Communications of the ACM*, 21(7).