Putting development and evaluation of core technology first

Anja Belz
NLTG, CMIS, University of Brighton, UK
A.S.Belz@itri.brighton.ac.uk

NLG needs comparative evaluation

NLG has strong evaluation traditions, in particular in user evaluations of NLG-based application systems (e.g. M-PIRO, COMIC, SUMTIME), but also in embedded evaluation of NLG components vs. non-NLG baselines (e.g. DIAG, ILEX, TAS) or different versions of the same component (e.g. SPOT). Recently, automatic evaluation against reference texts has appeared too, especially in surface realisation.

What has been missing are comparative evaluation results for comparable but independently developed NLG systems. Right now, there are only two sets of such results (for the SUMTIME weather forecasts, and for regenerating the Wall Street Journal Corpus). As a result, we have no idea at present what NLG techniques generally work better than others.

If NLG is a field of research that can progress collectively, rather than a loose collection of groups each progressing more or less independently, then it needs to develop the ability to comparatively evaluate NLG technology. This seems to me an absolutely fundamental principle for any branch of science and technology: without the ability to compare, results cannot be consolidated and there is no collective progress (Spärck Jones, 1981).

Shared tasks, but not necessarily shared data

That comparable techniques, components and systems need to perform comparable tasks — that comparative evaluation needs to be in that sense based on shared tasks — goes almost without saying. However, such tasks can be more or less loosely defined: implicitly by a set of paired inputs and outputs, or explicitly by a set of specifications and input/output requirements. Comparability increases if systems take the same type of inputs, and evaluation can be performed on the basis of a set of test inputs. Test-set evaluation can be useful in research-oriented evaluation, where results need to be obtained quickly and cost-efficiently. However, for evaluation at the application level, especially if it is user-based, test-input evaluation is often not necessary.

Core technology first, applications second

The single biggest challenge for comparative NLG evaluation is identifying sharable tasks: this is problematic in a field where systems are rarely developed for the same domain, let alone with the same input and output requirements.

One possibility is to propose an application for NLG researchers to develop systems for. These could then be evaluated according to ISO 9126 and 14598 on software evaluation, and this would shed light on the real-world usefulness of the systems.

However, NLG is a varied field with many applications and it will be hard to choose one that is recognised by a large enough number of researchers as their task. Moreover, evaluation at the application level would necessarily include application-specific content-determination techniques, and results would therefore not automatically generalise beyond the application. It would also not shed light on the usefulness or otherwise of any component technology.

We need an approach that unifies NLG, not one that creates a new subfield specialising in the chosen application. We need to focus on what unites NLG not what diversifies it. The way to do this is in my view to focus on the development and evaluation of core technology that is potentially useful to all NLG and to utilise the commonalities that have already evolved, in particular the more generally agreed subtasks such as GRE, lexicalisation, content ordering, or even a larger component like surface realisation.

Focus on output evaluation

The evaluation criteria general to all software systems covered by ISO standards 9126 and 14598 of course also apply to evaluating NLG systems, but we

still need to decide how to evaluate the — necessarily domain-specific — goodness of their outputs (one of the ISO criteria), and that is what research needs to focus on. Depending on how a shared task has been defined and whether a system or component is being evaluated, output evaluation could be in the form of added-value evaluation of components embedded within applications, direct evaluation of outputs or indirect evaluation by comparison against a set of reference texts. In terms of evaluation criteria, in the neighbouring disciplines of MT and summarisation, fluency and accuracy have emerged as standard criteria, and the latter now also assesses 'responsiveness' of a summary to the given topic, a criterion approximating 'real-world usefulness'.

Towards common subtasks, corpora and evaluation techniques

There are some subfields that have developed enough common ground to make it feasible to create a shared task specification straight away and have enough researchers able to participate (e.g. GRE). However, there is a lot that needs to be done to make this possible across larger parts of NLG.

Subtasks and input/output requirements need to be standardised to make core technologies truly comparable (as well as potentially reusable). In other NLP fields standardisation is often driven by evaluation efforts (e.g. in parsing), but it is probably more productive to work towards this in dedicated research projects. E.g. in the newly funded Prodigy Project, one of our core aims is to develop an approach to content representation that generalises to five different data-to-text domains.

Building data resources of NLG inputs and/or outputs may be the most straightforward way to encourage researchers to create comparable NLG systems. There are very few such resources at the moment, among them are the SumTime corpus, and the GREC corpus of short encyclopaedic texts for generating referring expressions in context that we are currently developing (Belz and Varges, 2007).

Creating NLG-specific evaluation techniques and assessing their reliability is essential so that we know how to reliably evaluate NLG technology. Such techniques should a assess the three criteria mentioned above: (i) language quality; (ii) appropriateness of content; and (iii) task-effectiveness, or how well do the generated texts achieve their communicative purpose.

We need a range of evaluation methods suitable for quick low-cost evaluation during testing of new ideas as well as reliable, potentially time and cost-intensive methods for evaluating complete systems. The aim of the GENEVAL initiative (Reiter and Belz, 2006) is to develop a range of evaluation techniques for NLG and to assess their reliability, ultimately aiming to provide NLG researchers with knowledge to decide which technique to use given their available time, resources and evaluative aim.

Concluding remarks

Comparative evaluation doesn't have to be in the shape of competitions with associated events (as opposed to just creating resources and encouraging other researchers to use them), but I happen to like the buzz and energy they create, the way they draw new people in, and the hot-housing of solutions they foster (Belz and Kilgarriff, 2006). It should at least be tried out to see whether it can work for NLG.

There's a lot of virtue in talking: discussing the options and trying to find consensus. But there's also virtue in doing — creating data and tasks and putting them out there for researchers to use if they want. Even organising competetive events to see if they work. The risks of getting it wrong seem small to me — shared-task evaluations can be run on a shoe-string (as SENSEVAL and CONLL continue to demonstrate), and anyway, these things have a habit of self-regulating: if an event, task or corpus fails to inspire people, it tends to quietly go away.

References

A. Belz and A. Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proc. INLG'06*, pages 133–135.

A. Belz and S. Varges. 2007. The GREC corpus: Main subject reference in context. Technical Report NLTG-07-01, Natural Language Technology Group, CMIS, University of Brighton.

E. Reiter and A. Belz. 2006. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proceedings of INLG'06*, pages 136–138.

K. Spärck Jones, 1981. *Information Retrieval Experiment*, chapter 12, page 245. Butterworth & Company.