

# Shared-task Evaluations in HLT: Lessons for NLG

Anja Belz

University of Brighton, UK

A.S.Belz@brighton.ac.uk

Adam Kilgarriff

Lexical Computing Ltd., UK

adam@lexmasterclass.com

## 1 Introduction

While natural language generation (NLG) has a strong evaluation tradition, in particular in user-based and task-oriented evaluation, it has never evaluated different approaches and techniques by comparing their performance on the same tasks (shared-task evaluation, STE). NLG is characterised by a lack of consolidation of results, and by isolation from the rest of NLP where STE is now standard. It is, moreover, a shrinking field (state-of-the-art MT and summarisation no longer perform generation as a subtask) which lacks the kind of funding and participation that natural language understanding (NLU) has attracted.

Evidence from other NLP fields shows that STE campaigns (STECs) can lead to rapid technological progress and substantially increased participation. The past year has seen a groundswell of interest in comparative evaluation among NLG researchers, the first comparative results are being reported (Belz and Reiter, 2006), and the move towards some form of comparative evaluation seems inevitable. In this paper we look at how two decades of NLP STECs might help us decide how best to make this move.

## 2 Shared-task evaluation in HLT

Over the past twenty years, virtually every field of research in human language technology (HLT) has introduced STECs. A small selection is presented in the table below<sup>1</sup>. NLG researchers have tended to be somewhat unconvinced of the benefits of comparative evaluation in general, and the kind of competitive, numbers-driven STECs that have been typical of NLU in particular. Yet STECs do not have to be hugely competitive events fixated on one task with associated input/output data and single evaluation metric, static over time.

**Tasks:** There is a distinction between (i) evaluations designed to help potential users to decide

whether the technology will be valuable to them, and (ii) evaluations designed to help system developers improve the core technology (Spärck Jones and Galliers, 1996). In the former, the application context is a critical variable in the task definition; in the latter it is fixed. Developer-oriented evaluation promotes focus on the task in isolation, but if the context is fixed badly, or if the outside world changes but the evaluation does not, then it becomes irrelevant. NLP STECs have so far focused on developer-oriented evaluation, but there are increasing calls for more ‘embedded’, more task-based types of evaluations<sup>2</sup>.

Existing NLP STECs show that tasks need to be broadly based and continuously evolving. To begin with, the task needs to be simple, easy to understand and easy for people to recognise as their task. Over time, as the limitations of the simple task are noted and a more substantial community is ‘on board’, tasks can multiply, diversify and become more sophisticated. This is something that TREC has been good at (still going strong 14 years on), and the parsing community has failed to achieve (see notes in table).

**Evaluation:** NLP STECs have tended to use automatic evaluations because of their speed and reproducibility, but some have used human evaluators, in particular in fields where language is generated (MT, summarisation, speech synthesis).

Evaluation scores are not independent of the task and context for which they are calculated. This is clearly true of human-based evaluation, but even scores by a simple metric like word error rate in speech recognition are not comparable unless certain parameters are the same: background-noise, language, whether or not speech is controlled. Development of evaluation methods and benchmark tasks therefore must go hand in hand.

Evaluation methods have to be accepted by the research community as providing a true approxi-

<sup>1</sup>Apologies for omissions, and for bias towards English.

<sup>2</sup>A prominent theme at the 2005 ELRA/ELDA Workshop on HLT Evaluation.

Name	Start	Domain	Sponsors	Notes
MUC	1987	Information extraction	US Govt	Rapidly came to define IE; ended 1998.
PARSEVAL	1991	Parsing	—	Only ever defined a metric, no STEC <sup>1</sup> .
TREC	1992	Information retrieval	US Govt	Large and long-running, multiple tracks.
SEMEVAL	1994	Semantic interpretation	US Govt	No STEC emerged <sup>2</sup> .
NIST-MT	1994	Machine translation	US Govt	Revitalised since 2001 by BLEU <sup>3</sup> .
Morpholympics	1994	Morphological analysis	GLDV	German morphological analysis; one-off.
SENSEVAL	1998	Word sense disambiguation	ACL-SIGLEX	Validity of WSD task problematic.
SUMMAC	1998	Text summarization	US Govt	One-off.
CoNLL	1999	Various	ACL-SIGNLL	Focus on learning algorithms.
CLEF	2000	IR across languages	EU Project	
DUC	2001	Document summarization	US Govt	Defines field.
Morfolimpiadas	2003	Morphological analysis	Portuguese Govt	Portuguese language only.
SIGHAN	2003	Chinese tokenization	ACL-SIGHAN	Key benchmark.
Blizzard	2003	Speech synthesis	Festvox project	Building synthetic voice from given data.
HAREM	2005	Named-entity recognition	Portuguese Govt	Portuguese language only.
RTE	2005	Textual entailment	EU Project	
TC-STAR	2005	Speech-to-speech translation	EU integrated project	Black-box and glass-box evaluation <sup>4</sup> .

### Notes

1. PARSEVAL is an evaluation measure, not a full STEC. This has proved problematic: the parsing community no longer accepts the PARSEVAL measure, but there has been no organisational framework for establishing an alternative.
2. SEMEVAL did not proceed largely because it was too ambitious and agreement between people with different interests and theoretical positions was not achieved. It was eventually reduced in scope and aspects became incorporated in MUC, SUMMAC and SENSEVAL.
3. MT has been transformed by corpus methods, which have shifted MT from a backwater to perhaps the most vibrant area of NLP in the last five years.
4. In TC-STAR, the SST task is broken down into numerous subtasks. The modules and systems that meet the given criteria are exchanged among the participants, lowering the barrier to entry.

mation of quality. E.g. BLEU is strongly disliked in the non-statistical part of the MT community because it is biased in favour of statistical MT systems. PARSEVAL stopped being used when the parsing community moved towards dependency parsing and related approaches.

**Sharing:** As PARSEVAL shows, measures and resources alone are not enough. Also required are (i) an event (or better, cycle of events) so people can attend and feel part of a community; (ii) a forum for reviewing task definitions and evaluation methods; (iii) a committee which ‘owns’ the STEC, and organises the next campaign.

Funding is usually needed for gold-standard corpus creation but rarely for anything else (Kilgarriff, 2003). Participants can be expected to cover the cost of system development and workshop attendance. A funded project is best seen as supporting and enabling the STEC (especially during the early stages) rather than being it.

In sum, STECs are good for community building. They produce energy (as we saw when the possibility was raised for NLG at UCNLG’05 and ENLG’05) which can lead to rapid scientific and technological progress. They make the field look like a game and draw people in.

### 3 Towards an NLG STEC

In 1981, Spärck Jones wrote that IR lacked consolidation and the ability to build new work on old, and that this was substantially because there was no commonly agreed framework for describing and evaluating systems (Spärck Jones, 1981, p. 245). Since 1981, various NLP sub-disciplines have consolidated results and progressed collectively through STECs, and have seen successful commercial deployment of NLP technology (e.g. speech recognition software, document retrieval and dialogue systems).

However, Spärck Jones’s 1981 analysis could be said to still hold of NLG today. There has been little consolidation of results or collective progress, and there still is virtually no commercial deployment of NLG systems or components.

We believe that comparative evaluation is key if NLG is to consolidate and progress collectively. Conforming to the evaluation paradigm now common to the rest of NLP will also help re-integration, and open up the field to new researchers.

**Tasks:** In defining sharable tasks with associated data resources for NLG, the core problem is deciding what inputs should look like. There is a real risk that agreement cannot be achieved on

this, so not many groups participate, or the plan never reaches fruition (as happened in SEMEVAL).

There are, however, ways in which this problem can be circumvented. One is to use a more abstract task specification describing system functionality, so that participants can use their own inputs, and systems are compared in task-based evaluations similar to the traditions and standards of software evaluation (as in Morpholympics). An alternative is to approach the issue through tasks with inputs and outputs that ‘occur naturally’, so that participants can use their own NLG-specific representations. Examples include data-to-text mappings where e.g. time-series data or a data repository are mapped to fault reports, forecasts, etc.

Both data-independent task definitions and tasks with naturally occurring data have promise, but we propose the second as the simpler, easier to organise solution, at least initially. A specific proposal of a set of tasks can be found elsewhere in this volume (Reiter and Belz, 2006). An interesting idea (recommended by ELRA/ELDA) is to break down the input-output mapping into stages (as in the TC-STAR workshops, see table) and then, in a second round of evaluations, to make available intermediate representations from the most successful systems from the first round. In this way, standardised representations might develop almost as a side-effect of STECs.

**Evaluation:** As in MT there are at least two criteria of quality for NLG systems: language quality (fluency in MT) and correctness of content (adequacy in MT). In NLG, these have mostly been evaluated directly using human scores or preference judgments, although recently automatic metrics such as BLEU have been used. They have also been evaluated indirectly, e.g. by measuring reading speeds and manual post-processing<sup>3</sup>. A more user-oriented type of evaluation has been to assess real-world usefulness, in other words, whether the generated texts achieve their purpose (e.g. whether users learn more with NLG techniques than with cheaper alternatives<sup>4</sup>).

The majority of NLP STECs have used automatic evaluation methods, and the ability to produce results ‘at the push of a button’, quickly and reproducibly, is ideal in the context of STECs. However, existing metrics are unlikely to be suitable for NLG

(Belz and Reiter, 2006), and there is a lot of scepticism among NLG researchers regarding automatic evaluation. We believe that NLG should develop its own automatic metrics (development of such metrics is part of the proposal by Reiter and Belz, this volume), but for the time being an NLG STEC needs to involve human-based evaluations of the intrinsic as well as extrinsic type.

**Sharing:** A recent survey conducted on the main NLG and corpus-based NLP mailing lists<sup>5</sup> revealed that there are virtually no data resources that could be directly used in shared tasks. Considerable investment has to go into developing such resources, and direct funding is necessary. This points to a funded project, but we recommend direct involvement of the NLG community and SIGGEN. Other aspects of organisation are not NLG-specific, so the general recommendations in the preceding section apply.

## 4 Conclusion

STECs have been remarkable stimulants to progress in other areas of HLT, through their community-building role, and through ‘hot-housing’ solutions to specific problems. There are also lessons to be learnt about STECs not being overly ambitious, remaining responsive to developments in the broader field and wider world, and having appropriate institutional standing. We believe that NLG can benefit greatly from the introduction of shared tasks, provided that an inclusive and flexible approach is taken which is informed by the specific requirements of NLG.

## References

- A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proc. EACL’06*, pages 313–320.
- A. Kilgarriff. 2003. No-bureaucracy evaluation. In *Proc. Workshop on Evaluation Initiatives in NLP, EACL’03*.
- E. Reiter and A. Belz. 2006. GENEVAL: A proposal for shared-task evaluation in NLG. In *Proc. INLG’06*.
- K. Spärck Jones and J. R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag.
- K. Spärck Jones, 1981. *Information Retrieval Experimentation*, chapter 12. Butterworth & Co.

<sup>3</sup>E.g. in the SkillSum and SumTime projects at Aberdeen.

<sup>4</sup>E.g. evaluation of the NL interface of the DIAG intelligent tutoring system, di Eugenio et al.

<sup>5</sup>Belz, March 2006.