

GENEVAL: A Proposal for Shared-task Evaluation in NLG

Ehud Reiter

University of Aberdeen, UK
ereiter@csd.abdn.ac.uk

Anja Belz

University of Brighton, UK
a.s.belz@brighton.ac.uk

Abstract

We propose to organise a series of shared-task NLG events, where participants are asked to build systems with similar input/output functionalities, and these systems are evaluated with a range of different evaluation techniques. The main purpose of these events is to allow us to compare different evaluation techniques, by correlating the results of different evaluations on the systems entered in the events.

1 Background

Evaluation is becoming increasingly important in Natural Language Generation (NLG), as in most other areas of Natural Language Processing (NLP). NLG systems can be evaluated in many different ways, with different associated resource requirements. For example, a large-scale task-effectiveness study with human subjects could last over a year and cost more than US\$100,000 (Reiter et al., 2003); on the other hand, a small-scale comparison of generated texts to human-written reference texts can be done in a manner of days. However, while the latter kind of study is very appealing in terms of cost and time, and cheap and reliable evaluation techniques would be very useful for people developing and testing new NLG techniques, it is only worth doing if we have reason to believe that its results tell us something about how useful the generated texts are to real human users. It is not obvious that this is the case (Reiter and Sripada, 2002).

Perhaps the best way to study the reliability of different evaluation techniques, and more generally to develop a better empirical understanding of the strengths and problems of different evaluation techniques, is to perform studies where a range of different evaluation techniques are used to evaluate a set of NLG systems with similar functionalities. Correlating the results of the different evaluation techniques will give us empirical insight as

to how well these techniques work in practice.

Unfortunately, few such studies have been carried out, perhaps because (to date) few NLG systems have been built with comparable functionality (our own work in this area is discussed below). We hope to surmount this problem, by organising ‘shared task’ events to which NLG researchers can submit systems based on a supplied data set of inputs and (human-written) text outputs. We will then carry out our evaluation experiments on the submitted systems. We hope that such shared-task events will also make it easier for new researchers to get involved in NLG, by providing data sets and an evaluation framework.

2 Comparative Evaluations in NLG

There is a long history of shared task initiatives in NLP, of which the best known is perhaps MUC (Hirschman, 1998); others include TREC, PARSEVAL, SENSEVAL, and the range of shared tasks organised by CONLL. Such exercises are now common in most areas of NLP, and have had a major impact on many areas, including machine translation and information extraction (see discussion of history of shared-task initiatives and their impact in Belz and Kilgarriff (2006)).

One of the best-known comparative studies of evaluation techniques was by Papineni et al. (2002) who proposed the BLEU metric for machine translation and showed that BLEU correlated well with human judgements when comparing several machine translation systems. Several other studies of this type have been carried out in the MT and Summarisation communities.

The first comparison of NLG evaluation techniques which we are aware of is by Bangalore et al. (2000). The authors manually created several variants of sentences from the Wall Street Journal, and evaluated these sentences using both human judgements and several corpus-based metrics. They used linear regression to suggest a combination of the corpus-based metrics which they be-

lieve is a better predictor of human judgements than any of the individual metrics.

In our work (Belz and Reiter, 2006), we used several different evaluation techniques (human and corpus-based) to evaluate the output of five NLG systems which generated wind descriptions for weather forecasts. We then analysed how well the corpus-based evaluations correlated with the human-based evaluations. Amongst other things, we concluded that BLEU-type metrics work reasonably well when comparing statistical NLG systems, but less well when comparing statistical NLG systems to knowledge-based NLG systems.

We worked in this domain because of the availability of the SumTime corpus (Sripada et al., 2003), which contains both numerical weather prediction data (i.e., inputs to NLG) and human written forecast texts (i.e., target outputs from NLG). We are not aware of any other NLG-related corpora which contain a large number of texts and corresponding input data sets, and are freely available to the research community.

3 Our Proposal

We intend to apply for funding for a three-year project to create more shared input/output data sets (we are focusing on data-to-text tasks for the reasons discussed in Belz and Kilgarriff (2006)), organise shared task workshops, and create and test a range of methods for evaluating submitted systems.

3.1 Step 1: Create data sets

We intend to create input/output data sets that contain the following types of representations:

- raw non-linguistic input data;
- structured content representations, roughly corresponding to document plans (Reiter and Dale, 2000);
- semantic-level representations, roughly corresponding to text specifications (Reiter and Dale, 2000);
- actual human-authored corpus texts.

The presence of intermediate representations in our data sets means that researchers who are just interested in document planning, microplanning, or surface realisation do not need to build complete NLG systems in order to participate.

We will create the semantic-level representations by parsing the corpus texts, probably using a LinGO parser¹. We will create the content representations using application-specific analysis tools, similar to a tool we have already created for SumTime wind statements. The actual data sets we currently intend to create are as follows (see also summary in Table 1).

SumTime weather statements: These are brief statements which describe predicted precipitation and cloud over a forecast period. We will extract the texts (and the corresponding input data) from the existing SumTime corpus.

Statistics summaries: We will ask people (probably students) to write paragraph-length textual summaries of statistical data. The actual data will come from opinion polls or national statistics offices. The corpus will also include data about the authors (e.g., age, sex, domain expertise).

Nurses' reports: As part of a new project at Aberdeen, Babytalk², we will be acquiring a corpus of texts written by nurses to summarise the status of a baby in a neonatal intensive care unit, along with the raw data this is based on (sensor readings, records of actions taken such as giving medication).

3.2 Step 2: Organise workshops

The second step is to organise workshops. We intend to use a fairly standard organisation (Belz and Kilgarriff, 2006). We will release the data sets (but not the reference texts), give people six months to develop systems, and invite people who submit systems to a workshop. Participants can submit either complete data-to-text NLG systems, or components which just do document planning, microplanning, or realisation.

We are planning to increase the number and complexity of tasks from one round to the next, as this has been useful in other NLP evaluations (Belz and Kilgarriff, 2006); for example, we will add surface realisation as a separate task in round 2 and layout/structuring task in round 3.

We will carry out all evaluation activities (see below) ourselves, workshop participants will not be involved in this.

3.3 Step 3: Evaluation

The final step is to evaluate the systems and components submitted to the workshop. As the main

¹<http://lingo.stanford.edu/>

²<http://www.csd.abdn.ac.uk/research/babytalk/>

Corpus	num texts	num ref (*)	text size	main NLG challenges
Weather statements	3000	300	1-2 sentences	content det, lex choice, aggregation
Statistical summaries	1000	100	paragraph	above plus surface realisation
Nurses' reports	200	50	several paras	above plus text structuring/layout

(*) In addition to the main corpus, we will also gather texts which will be used as reference texts for corpus-based evaluations; 'num ref' is the number of such texts. These texts will not be released.

Table 1: Planned GENEVAL data sets.

purpose of this whole exercise is to see how well different evaluation techniques correlate with each other, we plan to carry out a range of different evaluations, including the following.

Corpus-based evaluations: We will develop new, linguistically grounded evaluation metrics, and compare these to existing metrics including BLEU, NIST, and string-edit distance. We will also investigate how sensitive different metrics are to size and make-up of the reference corpus.

Human-based preference judgements: We will investigate different experimental designs and methods for overcoming respondent bias (e.g. what is known as 'central tendency bias', where some respondents avoid judgements at either end of a scale). As we showed previously (Belz and Reiter, 2006) that there are significant inter-subject differences in ratings, one thing we want to determine is how many subjects are needed to get reliable and reproducible results.

Task performance. This depends on the domain, but e.g. in the nurse-report domain we could use the methodology of (Law et al., 2005), who showed medical professionals the texts, asked them to make a treatment decision, and then rated the correctness of the suggested treatments.

As well as recommendations about the appropriateness of existing evaluation techniques, we hope the above experiments will allow us to suggest new evaluation techniques for NLG.

4 Next Steps

At this point, we encourage NLG researchers to give us their views regarding our plans for the organisation of GENEVAL, the data and evaluation methods we are planning to use, to suggest additional data sets or evaluation techniques, and especially to let us know whether they would be interested in participating.

If our proposal is successful, we hope that the project will start in summer 2007, with the first data set released in late 2007 and the first work-

shop in summer 2008. ELRA/ELDA have also already agreed to help us with this work, contributing human and data resources.

References

- Srinavas Bangalore, Owen Rambow, and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of INLG-2000*, pages 1–8.
- Anja Belz and Adam Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of INLG-2006*.
- Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL-2006*, pages 313–320.
- Lynette Hirschman. 1998. The evolution of evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12:283–285.
- Anna Law, Yvonne Freer, Jim Hunter, Robert Logie, Neil McIntosh, and John Quinn. 2005. Generating textual summaries of graphical time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL-2002*, pages 311–318.
- Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.
- Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of INLG-2002*, pages 97–104.
- Ehud Reiter, Roma Robertson, and Liesl Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.
- Somayajulu Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003. Exploiting a parallel text-data corpus. In *Proceedings of Corpus Linguistics 2003*, pages 734–743.